ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1682

Speech Emotion Recognition Using Voiced Segment Selection Algorithm

Yu Gu[1] and **Eric Postma** [2] and **Hai-Xiang Lin**[3] and **Jaap van den Herik** [4]

Abstract. Speech emotion recognition (SER) poses one of the major challenges in human-machine interaction. We propose a new algorithm, the Voiced Segment Selection (VSS) algorithm, which can produce an accurate segmentation of speech signals. The VSS algorithm deals with the voiced signal segment as the texture image processing feature which is different from the traditional method. It uses the Log-Gabor filters to extract the voiced and unvoiced features from spectrogram to make the classification. The finding shows that the VSS method is a more accurate algorithm for voiced segment detection. Therefore, it has potential to improve performance of emotion recognition from speech.

1 Introduction

A speech always consists of voiced and unvoiced parts. When vocal chords vibrate to pronounce vowels, voiced segments are generated. In contrast to unvoiced segments, voiced segments show periodic and prosodic signals. Unvoiced segments show irregular signals, generated by the influence of narrow vocal tract. Emotion is an important component of the information contained in a speech. Emotional information in speech is represented in a variety of prosodic types and is also mainly contained in the voiced parts. That is why researchers always focus on the voiced parts of a speech in emotion recognition.

The outline of the paper is as follows. Section 2 describes the methodology of VSS method in detail. In Section 3, the set-up of the comparative evaluation of the VSS method is described. Experimental results and discussions are presented in Section 4. Finally, Section 5 provides concluding remarks.

2 Formulation of VSS Method

There are two key components in the VSS method which are the spectrogram and the log-Gabor filter.

The VSS method deals with the voiced activity detection as a classification issue. In contrast to existing machine learning algorithm for VAD which normally used the acoustic features or statistical features for the classification, we propose a novel kind of feature which is extracted by log-Gabor filter from the spectrogarm for this issue. We use two-dimensional Gabor filters which are locally corresponding to the orientations of energy bands in the spectrogram. Therefore, by convolving the spectrogram with a Gabor filter of a given spatial frequency and orientation, the convolved spectrogram represents spectro-temporal patterns with the associated spatial frequency (width) and orientation, respectively. We use the log-Gabor filters with combination of scale and orientation for convolving whole spectrogram to obtain the texture information. Because all the texture patterns are generated by voiced parts in the spectrogram, these convolution results from the filters are sensitive to the voiced segment for speech signal. Each combination of different scale and orientation will result in one filter-image. The energy of each filter-iamge will be averaged by the time sequence and be stored as a vector data. And finally, all the vectors which calculated by each filters will be assemble as a matrix. The Support vector machine will be used to classify the voiced and unvoiced segment by using the convolution feature matrix from spectrogram.

3 Experiment of the VSS method

This section describes the implementation which is used in the experiments and evaluation details.

The evaluation of the VSS method was performed on two corpora: the Mandarin Affective Speech (MAS) corpus (MAS, 2007) and the Berlin Database of Emotional Speech (Emo-DB, 2007). More details about the corpus can be found via the Linguistic Data Consortium website[5].

The experimental procedure consisted of three steps: (i) spectrogram generation, (ii) spectrogram convolution using Log-Gabor filters, (iii) voiced and unvoiced segment classification. In the following, the details of each of these steps will be outlined.

Table 1. The value of parameters for Log-Gabor filter

name of parameter	value of parameter
spectrogram time resolution t	0
scale N_s and orientation N_o	12 12
segmented Patch	8
minwavelength	3
mult	1.35
sigmaOnf	0.8

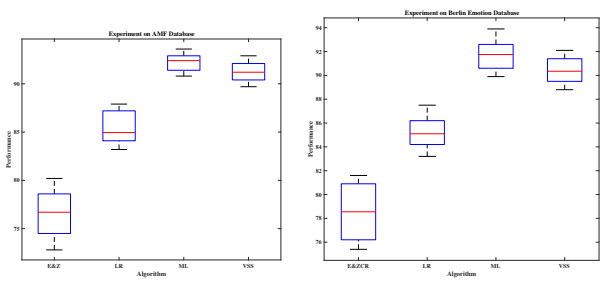
(i) The speech signal was visualized in spectrogram. Each auditory signal (utterance) was transformed into a spectrogram using Matlab's spectral analysis function. (ii) Each spectrogram is convolved by a Log-Gabor filter with a number different parameter values of scale

[1] Tilburg center of Communication and cognition, Tilburg University, Tilburg, Netherlands, email: y.gu_1@uvt.nl
[2] Tilburg center of Communication and cognition, Tilburg University, Tilburg, Netherlands
[3] Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands
[4] Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands

[5] http://catalog.ldc.upenn.edu/LDC2007S09

and orientation. N_s and N_o denote the scale and the orientation respectively. Therefore, if the number of scales and the orientations are equal to N_s and N_0 respectively, then, $N_s * N_o$ bank filter images are generated. The parameters of log-Gabor filters as shown in Table 1 will be applied to make the convolution to the spectrogram. The convolution values within each image is averaged by the time sequence and store as a vector. Thus, each vector will have $1 * 512$ value. (iii) Step (ii) will produce $N_s * N_o$ convolution image, and the total number of the speech record is N_t. Therefore $N_s * N_o * N_t$ vectors which we could achieve after the whole convolution. Combining the energy value vectors into matrices features which we can use for training an SVM classifier.

The performance evaluation consists of two parts: a comparison in accuracy of voiced part detection between the VSS method exiting voiced detection methods and the performance in emotion recognition by the VSS method. First, we compare the VSS method with another prevalent VAD method in the voiced segment detection. Three major algorithms were duplicated: the a) The Energy and ZRC method, b) the statistical likelihood ratio method [1] and c) the Deep belief network method [2].



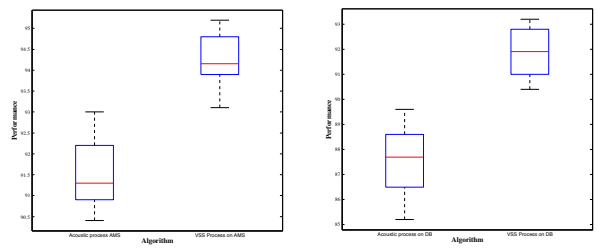
(a) AMS database (b) Berlin database

Figure 1. Comparison of the voiced part accuracy performances obtained on both database.

Figure 1 show the comparison of the voice accuracy between the three major methods and our VSS method by using the AMS and DB database respectively. We can clearly observe that our method was significantly better than the traditional ZCR and LR method. There are three possible explanations are as follows. Firstly, the spectrogram contains various kinds of most occurring acoustic phenomena, such as hamonicity, formants, vertical edges which are all the crucial characters for indicating the voice parts. And the formant and vertical edges are exactly the start and end point for the voice signal. Therefore, it can detect all acoustic phenomena which in other words to recognize the voice speech in a robust manner. Secondly, traditional methods have limitation due to manually set thresholds, the value of threshold was not always optimized. Compare to that, the SVM has a strong ability for the high dimensional feature learning that for the voice and unvoiced discrimination which can be less influence than manually and more accurate. Last reason for the higher accuracy is due to that, the log-Gabor filters has strong ability to distinguish the phonetic phenomena from noise background in spectrogram. Even through the noise in speech which can bring a huge number of corresponded peaks to the spectrogram. The reason which can be explained this is because the output of Log-Gabor filters is achieved from integrating the entire 2-D spectrogram information, which make the 2-D filterbank more robust ability to the noise. Moreover, the DBN method slightly outperforms the VSS method. However, for the deep learning it requires huge computation consumption. Therefore, we conclude that the VSS method is a useful and practical method to improve the voice detection accuracy than the traditional method.

4 Experiment of Speech Emotion Recognition

In the speech emotion recognition experiment, the acoustic features were extracted from the voiced segments based on the VSS method. The performance of speech emotion recognition was also compared with and without VSS procedure before the feature extraction. To avoid over-fitting due to the PCA and the SVM parameter optimization, the evaluation was performed using a nested 10×10-fold cross validation,.



(a) AMS database (b) Berlin database

Figure 2. Comparison of the recognition performances obtained from both database.

The classification was applied in both corpus of MAS and Emo-DB. Figure 2 visualizes the classification performances with and without VSS before feature extraction. Both of the features were base on the state-on-art baseline. The comparative evaluation of employing VSS method against without VSS method. It demonstrates that the performance of speech emotion recognition with VSS method led to a non-overlapping improvement. The results showed that VSS method for voiced detection is more effective in extracting features, because most of the acoustic features are related to the voiced parts.

5 Conclusions

We proposed a novel VSS algorithm which uses the log-Gabor filter to extract the features from the spectrogram for the voiced segment classification. Experiment results showed that the VSS algorithm can be a useful and practical method for the voiced activity detection. Furthermore, the performance in speech emotion recognition shows that the classification rate is improved for both Chinese and German speech. It convinced us the VSS is an useful compliment for the speech emotion recognition.

REFERENCES

[1] Youngjoo Suh and Hoirin Kim, 'Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection.', *IEEE Signal Process. Lett.*, **19**(8), 507–510, (2012).

[2] Xiao-Lei Zhang and Ji Wu, 'Deep belief networks based voice activity detection', *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**(4), 697–710, (2013).