



ISSN: 0975-833X

RESEARCH ARTICLE

A NOVEL KEYWORD SPOTTING APPROACH IN SPEECH MINING USING WAVELET PACKET TRANSFORMATION

***Senthil Devi, K. A. and Dr. Srinivasan, B.**

Assistant Professor of Computer Science, Gobi Arts and Science College, Gobichettipalayam

ARTICLE INFO

Article History:

Received 26th May, 2016
Received in revised form
20th June, 2016
Accepted 27th July, 2016
Published online 31st August, 2016

Key words:

Speech Mining, Keyword Spotting,
MFCC, Discrete Wavelet Transform,
Wavelet Packet Decomposition,
Euclidean Distance.

ABSTRACT

Keyword spotting (KWS) is an innovative research area and has many applications in speech mining. Keyword spotting is the task of identifying the occurrences of certain desired keywords in an arbitrary speech signal. In this paper, we propose a new keyword spotting approach using Wavelet Packet Decomposition (WPD) and sliding frames methodology (WPDSFM). The proposed method is capable of identifying a target keyword in a continuous speech of any length. Euclidean distance is calculated between keyword and a block of speech frames for similarity measure. Experiments are done to compare that the result of the proposed WPDSFM approach with the results of MFCC and Discrete Wavelet transform (DWT) features based sliding frame methods. The performance measures show that the proposed approach is better than that of the other two approaches in terms of accuracy.

Copyright©2016, Senthil Devi and Dr. Srinivasan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Senthil Devi, K.A. and Dr. Srinivasan, B. 2016. "A novel keyword spotting approach in speech mining using wavelet packet transformation", *International Journal of Current Research*, 8, (08), 36943-36946.

INTRODUCTION

Speech Mining is a research field that aims to bridge the gap between speech processing and text data mining methods. Keyword spotting is an excellent technology in speech mining (Thambiratnam and Albert, 2004). It is a retrieval of all instances of a given keyword in spoken utterances. KWS generally carries some kind of classification based upon speech features which are usually obtained via time-frequency representations. It eliminates a lot of human work in such tasks like audio data mining, named entity search, state security domain, mobile applications, weather forecasting, agriculture, audio indexing, keyword monitoring, healthcare, automatic translation, robotics, telephone routing, video games, transcription etc. Keyword spotting methods are classified into three major categories: LVCSR-based, phone-lattice based and acoustic keyword spotting (Bridle, 1973). In the task of acoustic keyword spotting, it's unnecessary to recognize the whole sentence of the utterance. The performance of acoustic keyword spotting system could be running far more quickly in real time because of the no considerations on a large language model, so that it's chosen as our system.

A variety of strategies have been proposed to improve acoustic keyword spotting. One of the keyword spotting strategies is sliding frame-based method. The method was first proposed in (Silaghi and Bourlard, 2000) which used sliding a frame-based keyword template along the speech signal and using a nonlinear dynamic time warping algorithm to efficiently search for a match. This sliding frame model was also proposed with other approaches in (Wilpon *et al.*, 1989) and (Zhang and Glass, 2009). In (Bahi and Benati, 2009), audio keyword spotting was addressed with Gaussian Mixture Model (GMM) which was trained to represent each speech frame with a Gaussian posteriorgram. A segmental dynamic time warping (SDTW) technique is used to compare the Gaussian posteriorgrams between predefined keyword and unseen test data. Keyword spotting was achieved by processing the distortion scores on the test utterances. Another keyword spotting approach presented in (Khan and Holton, 2014) used MFCC and energy of the speech signal as feature set. This system is based on Vector Quantitation (VQ) and Hidden Markov Model (HMM) that need a high computational time for training the learning models. Recently the sliding frame model was proposed in (Jothilakshmi, 2013). In this work, wavelet transform is applied to the enhanced speech and Euclidean distance is calculated between test and template word's features in sliding frame methodology.

***Corresponding author: Senthil Devi, K.A.**

Assistant Professor of Computer Science, Gobi Arts and Science College, Gobichettipalayam.

A combination of the sliding frame methodology and Neural Network was also proposed in (Sangeetha and Jothilakshmi, 2014). In this paper, the distribution capturing ability of the Auto Associative Neural Network (AANN) is used with MFCC features for spoken keyword detection. The sliding frame method with MFCC and Support Vector Machine (SVM) was proposed in (Davis and Mermelstein, 1980). The method used support vector machine (SVM) misclassification rates obtained from the hyperplane of two classes to efficiently search for a match. Speech signals possessing non-stationary character are not well suited for detection and classification by traditional Fourier methods. It has been shown that wavelets can approximate time varying non-stationary signals in a better way than the Fourier transform representing the signal on both time and frequency domains (Gokhale and Daljeet Kaur Khanduja, 2010). Furthermore, wavelet decomposition allows analyzing a signal at different resolution levels. Hence, in this paper, wavelet packet decomposition based sliding frame model is proposed.

Keyword Spotting System

Keyword Spotting is conceptually simpler alternative to speech recognition followed by text mining and is based on acoustic matching between the keyword and the speech signal. KWS aims to extract the partial information from speech signal in the form of a phrase or keyword. There are five different stages in spotting keyword which are shown in Fig. 1.

They are namely, creating the speech database, preprocessing, wavelet transformation, feature extraction and keyword detection. Among these stages, pre-processing and transformation of speech are considered as crucial steps to reduce noises and to develop a robust and efficient KWS system. The feature extraction stage is a key, because better feature is good for improving accuracy rate. The extracted features of both keyword and input speech are compared and similarity score is calculated between them. The number of occurrences of keyword in the input speech is counted based on the threshold on the similarity distance.

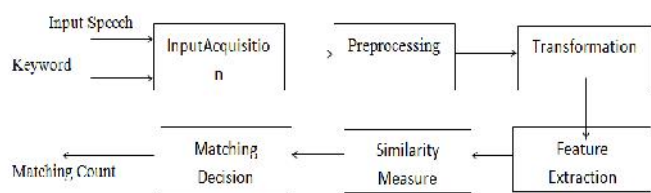


Figure 1. Basic components of keyword spotting system

Feature Extraction for keyword detection

The speech signal in its original form has great number of data points. A feature extraction method is required to choose the most useful information from the speech signal and keyword to shape a feature vector in a lower dimensionality, and eliminate any redundant and irrelevant information that may have disadvantageous effects on the keyword detection. Mel-frequency Cepstrum Coefficient (MFCC) and wavelet based feature extraction techniques are discussed in this section.

MFCC Feature Extraction

MFCC has proven to be one of the most successful feature representations in speech processing tasks. Fig.2. illustrates the working process of MFCC technique. Firstly, the input speech signal is segmented as frames of N samples each. Windowing technique is then applied to each individual frame in order to minimize the signal discontinuities from beginning until the end of each frame.

The frames are then converted from time-domain into frequency-domain via Fourier transform. The signal is converted next into spectrum form and followed by the mel-frequency wrapping process to acquire its mel spectrum. In this process, there are several important parameters that are needed to be determined such as frequency range limit, number of coefficients and number of filter banks. Finally, the obtained feature vectors will be averaged in order to acquire its distinctive features for training purposes.

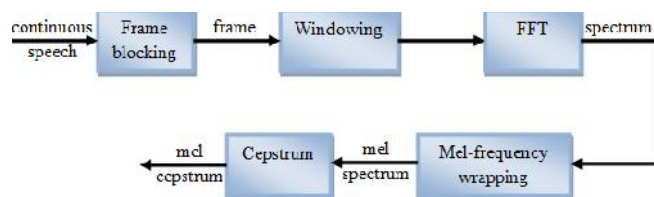


Fig. 2. Working process in MFCC computation

DWT based Feature Extraction

The DWT can be used for Multi Resolution Analysis (MRA). Since speech signal is a non-stationary and nonlinear signal MRA can be used. The given signal is decomposed into the approximation and detail coefficients. There are many types of wavelets such as Haar, Daubechies, Discrete Meyer, and others which can be used to perform wavelet transform. In addition, different levels of decomposition can be applied to speech signal in order to obtain desired output. At the first level of decomposition, a low-pass filter with corresponding wavelet is applied to the signal in order to obtain its approximation coefficients. For the next level of decomposition, the approximation coefficients from the previous level are again convoluted with the same type of wavelet. On the other hand, a high-pass filter with the same wavelet is applied to the desired level of approximation coefficients in order to obtain its detail coefficients for further processing.

WPD based Feature Extraction

A more comprehensive form of the standard wavelet transform is the wavelet packet, which decomposes both the high and low frequency bands at each level. A pair of low- and high-pass filters is used to recognize two sequences capturing dissimilar frequency sub-band features of the original signal. These sequences are then decimated. This process can be repeated to partition the frequency spectrum into smaller frequency bands for obtaining different features while detecting the temporal information. Wavelet packet atoms are waveforms indexed by three naturally interpreted parameters, position, scale, frequency.

WPT features have better presentation than the DWT. The wavelet packet decomposition is shown in the Fig. 3.

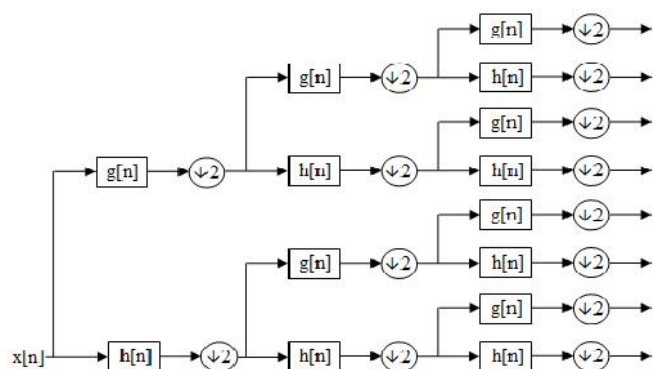


Figure 3. Wavelet Packet Decomposition Tree

Proposed Spoken Keyword Spotting (WPDSFM) Method

There are a number of approaches were designed for keyword spotting that use Fourier Transform for the spectral analysis. Most of the approaches used MFCC feature vectors for the similarity measure between keyword and template speech. W. Khan et al. (2013) used wavelet transform based features for spotting keyword. In the proposed WPDSFM method, combination of sequential processes is implemented. The input speech and keyword given are passed to pre-processing stage which enhances the speech quality. Then the wavelet packet decomposition is applied on the enhanced speech signals. WPD based features for every single frame of the given search keyword and speech signal are obtained. The extracted features for both test and template frames are passed to Euclidean distance to measure the similarity score. As Euclidean distance provides dissimilarity score, fewer score means more similar. Figure 4 shows the sequential flow of the all processes and detail of each component is addressed in the following section.

Preprocessing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. Speech signal pre-processing covers digital filtering, to enhance the speech quality in terms of silence removal, noise reduction, resampling and segmentation. In this proposed system moving – average filter function is used to filter the input speech and keyword given.

WPD based Feature Extraction

The wavelet packet decomposition is applied to the enhanced speech signal to acquire its frequency domain spectrum and filter out unwanted frequencies from input and template speech. The selected frequency spectrum is passed to feature extraction process that extracts some important features out of time and frequency domain speech signal. The features which are extracted and used for the test and template frame matching are listed below:

- RMS (Root Mean Square level)
- Correlation

- Homogeneity
- Standard Deviation
- Variance
- Smoothness
- Kurtosis
- Skewness

Keyword detection

A keyword is defined by an ordered sequence of acoustic elements. The proposed method tries to detect this sequence in the given audio stream by splitting the audio stream into blocks in the size of keyword. Sliding frame method is used. In this process, initially a block of frames such that the number of frames in the block is equal to number of frames of the keyword signal are selected from the input signal starting from the first frame. This block of feature vectors is used for finding similarity measure using Euclidean distance. If the word corresponding to the block of frames is same as the keyword then the score is the minimum. If the word corresponding to the block of frames is completely different from the keyword, the feature vectors from the block may not fall into the distribution and the similarity measure gives high score. Likewise, the next possibility is the word corresponding to the block of frames is partially similar to the keyword. If this is the case, the similarity score of the block will be in between the above two values. If the obtained score is within the limit of the specified threshold, then the matching count is incremented by one. After the above processing of the current block, the block is shifted by one frame to the right. Then the entire procedure is repeated for this new block and the similarity score is obtained. Likewise the scores are obtained and matching is counted until the tail end of the block reaches the last frame of the speech frames.

Proposed Keyword Spotting WPDSFM Algorithm

Consider the features of the input signal $S = \{S_i : i = 1, 2, 3, \dots, n\}$ where i is the frame index and n is the total number of frames in the input speech signal and the speech features of the keyword $K = \{K_j : j = 1, 2, 3, \dots, m\}$ where j is the frame index and m is the total number of frames in the search keyword signal. The proposed WPDSFM algorithm for retrieving the speech files for the given search speech keyword is summarized as follows:

- Consider a block of frames as X and l is the index of the speech block X and initially l is taken as 1.
- From frames of input speech signal, m number of frames are selected from l th position as a block $X_l = \{S_a : a = l+1, \dots, l+m\}$ with the constraint $l+m \leq n$.
- Euclidean distance is used to calculate the similarity measure (dis) in between the features of keyword K and features of a block of input speech X_l .
- The threshold value of similarity distance T is specified (T has 0.5 in our experiment). The similarity score dis is compared with the threshold value T .
- If the $dis \leq T$ then

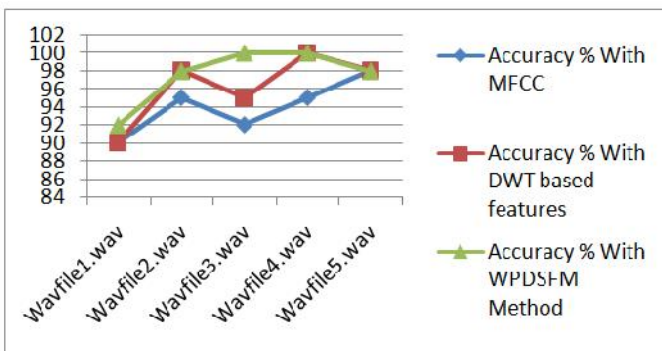
- Matching is found and the value of match count is incremented by 1.
- The value of l is added with m.
- Else
- No matching is found.
- The value of l is added with 1.
- The steps 2 to 5 are repeated until l+m exceeds n.

Table 1. Speech Database

Input spoken file	Keyword	Length
Wavfile1.wav	Ondru	Three word Utterance of 4secs length
Wavfile2.wav	Iradnu	Six word Utterance of 5secs length
Wavfile3.wav	Ondru	Eight word Utterance of 13secs length
Wavfile4.wav	Nadu	Nine word Utterance of 15secs length
Wavfile5.wav	Nandru	Ten word Utterance of 8 secs length

Table 2. Accuracy of MFCC, DWT and WPDSFM methods

File Name	Accuracy %		
	With MFCC	With DWT based features	With WPDSFM Method
Wavfile1.wav	90	90	92
Wavfile2.wav	95	98	98
Wavfile3.wav	92	95	100
Wavfile4.wav	95	100	100
Wavfile5.wav	98	98	98

**Figure 4. Comparison of results in MFCC, DWT and WPD methods**

Experimental Results

In the proposed WPDSFM Algorithm, Wavelet packet decomposition based features are used with sliding frame technique. A speech database is designed with 10 speakers. Speech was recorded in a silent room environment with a PC computer with AUDACITY sound recording package in frequency 8000Hz. For each speaker we recorded five utterances with different length and speech contents of Tamil language speech. Every speaker read the same sentences. Summary of the speech database is given in the table 1. The WPDSFM algorithm extracts a number of features from the pre-processed and wavelet packet decomposition based filtered signal that include RMS, Correlation, Homogeneity, Standard Deviation, Variance, Smoothness, Kurtosis, Skewness and pass these features for the similarity match analysis. The similarity score threshold value is taken as 0.5. Keyword spotting accuracy is calculated.

The method is repeated with MFCC and DWT based features in the same sliding frame method. The results are compared and it is shown in the Figure 4. The code is developed using MATLAB 2014a.

Conclusion

An approach for Keyword spotting in speech mining is designed. The approach uses wavelet packet decomposition and sliding frames methodology (WPDSFM) with Euclidean distance with the method is of identifying a target keyword in a continuous speech of any length. The approach is also compared with the results of MFCC and Wavelet transform features based sliding frame methods. The experimental results show that the proposed approach competes with the later methods in terms of accuracy and computational time.

REFERENCES

- Bahi, H., and Benati, N. 2009. A new keyword spotting approach. *IEEE International Conference on Multimedia Computing and Systems*, pp. 77-80.
- Bridle, J. S. 1973. "An efficient elastic-template method for detecting given words in running speech", *In Proc. of the Brit. Acoust. Soc. meeting*.
- Davis, S.B., Mermelstein, P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. Acoust. Speech Signal Process*, 1980, 357–366.
- Gokhale, M. Y. Daljeet Kaur Khanduja, 2010, "Time Domain Signal Analysis Using Wavelet Packet Decomposition Approach", *Int. J. Communications, Network and System Sciences*, 2010, 321 – 329.
- Jothilakshmi, S. 2013. Spoken keyword detection using auto associative neural networks, *International Journal Speech Technology*, Springer, 2013, pp. 83-89.
- Khan, W. and Holton, R. 2014. Word spotting in continuous speech using wavelet transform, *IEEE International Conference on Electro/Information Technology*, 2014, pp. 275-279.
- Sangeetha, J. and Jothilakshmi, S., "A novel spoken keyword spotting system using support vector machine", *Engineering Applications of Artificial Intelligence*, Springer, 2014, pp. 287–293.
- Silaghi, M. C., and Bourlard, H. 2000., "Iterative posterior-based keyword spotting without filler models", *In Proc. of ICASSP*.
- Thambiratnam, Albert J. K 2004. Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting", *IEEE Xplore*, 2004
- Wilpon, J. G., Rabiner, L. R., Lee, C. H., and Goldman, E. R. 1989. "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech", *In Proc. of ICASS*, 1989.
- Zhang, Y. and Glass, J. R. 2009. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams, *IEEE Proceedings of the Automatic Speech Recognition & Understanding (ASRU)*, 2009.
