Background Noise Identification System Based on Random Forest for Speech

Shambhu Shankar Bharti, Manish Gupta and Suneeta Agarwal

Abstract Background noise is acoustically added with human speech while communicating with others. Nowadays, many researchers are working on voice/speech activity detection (VAD) in noisy environment. VAD system segregates the frames containing human speech/only noise. Background noise identification has number of applications like speech enhancement, crime investigation. Using background noise identification system, one can identify possible location (street, train, airport, restaurant, babble, car, etc.) during communication. It is useful for security and intelligence personnel for responding quickly by identifying the location of crime. In this paper, using VAD G.729, a new algorithm is proposed for selecting an appropriate set of noisy frames. Mel-frequency cepstral coefficient (MFCC) and linear predictive coding (LPC) are used as feature vectors. These features of selected frames are calculated and passed to the classifier. Using proposed classifier, seven types of noises are classified. Experimentally, it is observed that MFCC is a more suitable feature vector for noise identification through random forest classifier. Here, by selecting appropriate noisy frames through proposed approach accuracy of random forest and SVM classifier increases up to 5 and 3%, respectively. The performance of the random forest classifier is found to be 11% higher than SVM classifier.

Keywords VAD · VAD G.729 · Random forest · SVM · Noise identification Noise extraction

S.S. Bharti (✉) · M. Gupta · S. Agarwal
MNNIT Allahabad, Allahabad, India
e-mail: shambhu4u08@gmail.com

M. Gupta
e-mail: manishymca2007@gmail.com

S. Agarwal
e-mail: suneeta@mnnit.ac.in

S.S. Dash et al. (eds.), *International Conference on Intelligent Computing and Applications*, Advances in Intelligent Systems and Computing 632,
https://doi.org/10.1007/978-981-10-5520-1_30

1 Introduction

Identification of background noise is a basic but tedious problem in audio signal processing. Till now, it has got less attention of the researchers working on audio signal processing. During communication through speech, background noise gets acoustically added with the speech signal. So, this noise signal is also communicated with the clean speech. These signals carry information about the background location (street, train, airport, restaurant, babble, car, etc.) of the person during communication. By identifying the type of background noise, one can easily identify the possible location of a person at the time of communication. For example, if a person is communicating using mobile phone, then the region of that person can be traced through mobile signal tower. Using background noise identification system, possible location within that region can be identified which will reduce the search space.

2 Previous Work

Chu et al. [1] used "composite of deep belief networks (composite-DBNs)" for recognizing 12 different types of common environmental sounds. Mel-frequency cepstral coefficient (MFCC) and matching pursuit (MP) are used as feature vectors. Maximum accuracy in environmental sound classification using composite-DBNs has been claimed as 79.6% by taking MFCC and MP features.

Frequency component of maximum harmonic weight (FCOMHW), local search tree, effective segment length of audio data (ESLOAD) and first-order difference mel-frequency cepstral coefficients matrix (D-MFCCM) features have been used by Li [2] to classify environmental sound.

Toyoda et al. [3] used combination of instantaneous spectrum at power peak and the power pattern in the time domain as features. Multi-layered perception neural system is used for the environmental sound classification. This classifier is used to classify 45 environmental sounds. Classification accuracy was about 92% claimed in the paper.

Pradeep et al. [4] used audio features like zero-crossing rate (ZCR), LPC, linear predictive cepstral coefficient (LPCC) and (log frequency cepstral coefficients) LFCC. Gaussian mixture model was used in the paper for modelling an event. Training audio data frame was taken of 50 ms. Using single Gaussian classifier, 76 and 80% accuracy for walking and running events, respectively, have been claimed.

Lozano et al. [5] used audio features like MFCCs, ZCR, centroid and roll-off point for audio classification. Maximum accuracy ratio by type of sound was about 81.42% claimed in this paper.

Han and Hwang [6] used traditional features (TFs) like ZCR, MFCC; change detection features (CDFs) like chirp rate spectrum; and acoustic texture features (ATFs) like discrete wavelet transform (DWT), discrete curvelet transform for

sound classification. Nonnegative matrix factorization [7] has been used for dimensionality reduction, and SVM has been taken as classifier. Maximum accuracy was claimed as 86.09%.

Kraetzer et al. [8] used data mining tool WEKA with *K*-means as a clustering and naive Bayes as a classification technique for the classification process. AAST (AMSL Audio Steganalysis Toolset, version 1.03) [9] and 56 mel-cepstral domain-based features are used.

Following are the main objectives of this paper:

(i) To develop noise identification system using random forest classifier.
(ii) To develop an algorithm for noise extraction using VAD G.729.
(iii) To find suitable feature/features for noise identification system.
(iv) To find suitable classifier for noise identification system.

The rest of this paper is organized as follows. In Sect. 3, random forest is discussed. Section 4 explains the proposed approach. In Sect. 5, parameters used for performance evaluation of the system are discussed. Section 6 explains the experimental set-up followed by summarization of the results. Section 7 concludes the paper.

3 Random Forest

Random forest or random decision forest [10, 11] uses an "ensemble learning method" for classification. It works by constructing a multitude of decision trees at training time. It identifies the class using mode of the classes or mean prediction (regression) of the individual trees. The general method of random decision forests was first proposed by Ho in 1995 [10]. Two well-known methods are boosting [12] and bagging [13] used in classification trees. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. For example, if there are n classes for classification, then samples of all the classes are selected randomly for the training purpose. If there are K input variables in each sample, then k $(k < K)$ is specified at each node where k variables are selected at random. Each tree grows at its maximum extent without any pruning. New data are classified considering maximum vote gained from different trees.

4 Proposed Approach

Noise extraction plays an important role in noise identification system. Noise can be better classified if the noisy frames are extracted correctly. In speech communication system, a frame either contains both human speech and noise or noise only.

In this paper, a new algorithm has been proposed for better noise identification by selecting the subset of noisy frames obtained from VAD G.729. Here, a framework is also proposed for noise identification. This framework may also be useful for language identification. Classifiers are trained using the feature/features (LPC, MFCC, LPC and MFCC together) extracted from Aurora2 noise.

Proposed Algorithm for Noise Extraction:

INPUT: Speech signal S of size $N1 \times 1$.

OUTPUT: Noise signal of size $N2 \times 1$.

Assumptions and Notations:

(i) S is a vector of size $N1 \times 1$. It contains $N1$ samples of speech signal.
(ii) N is the total number of frames.
(iii) $N2$ ($N2 <= N1$) is the number of samples in noise signal.
(iv) out is a matrix of size $N \times 2$. Matrix out will have two columns. First column represents frame number and second column represents *VAD* technique output.
(v) ||, represents the logical "***OR***" operation.
(vi) flag, flag1 and flag2 are binary variables.

Procedure noise_extract(S):
Repeat step 1 and step 2 for i = 1 to N.

Step 1: *Apply VAD G.729 technique on ith frame and store its output with corresponding frame number.*

// VAD G.729 output will be either 1 or 0. 1 for those frames that contain human speech and 0 for others.

Step 2: Save the result in matrix out.

Repeat step 3 and step 4 for i = 3 to N − 1.

Step 3: *If out(i, 2)* equals to *0*, then

flag1 = *out(i-1,2) || out(i-2,2).*
flag2 = *out(i-1,2) || out(i + 1,2).*
flag1 = *flag1 || flag2.*
end If

Step 4: *If flag* equals to *0*, then

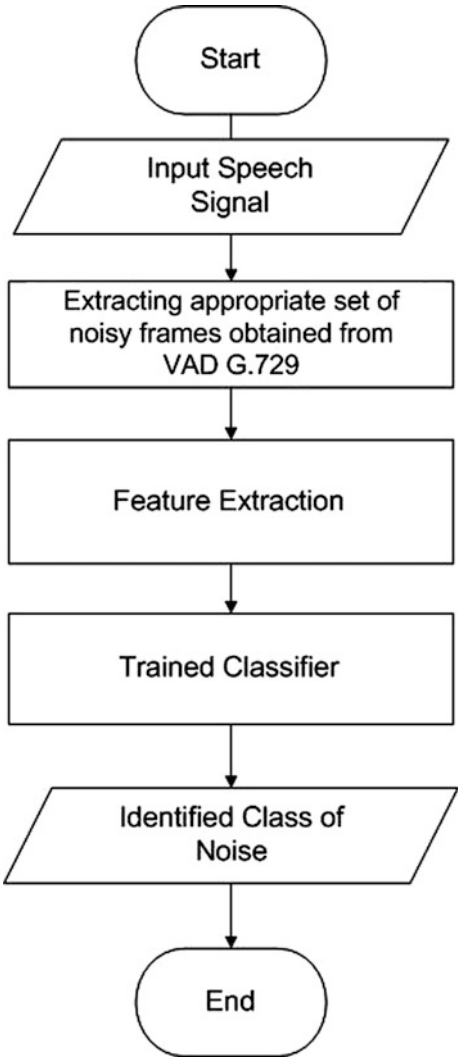
Add the samples of that frame into noise signal.
end If

Step 5: *Return noise_signal.*
End Procedure.

Feature vectors are calculated for noise_signal. Noise is identified by trained classifier using these features.

5 Parameters Used for Measuring the Performance

Reliability ratio (RRN), accuracy ratio (ARN) [5] and confusion matrix for the noise are used for measuring the performance of the proposed system. Suppose in a given signal S, noise of K different classes is present. TF_i is the total numbers of frames with noise of class i in the given signal. FC_i is the total number of frames having noise of type i as classified by classifier while T_i is the number of frames truly classified as of class i by the classifier.



Block diagram of proposed framework for noise identification system

Reliability ratio by type of noise (RRN)
It is formulated as:

$$\mathrm{RRN} = \frac{\sum_{i=1}^{k} \frac{T_i}{\mathrm{FC}_i}}{K} \tag{1}$$

Accuracy ratio by type of noise (ARN)
It is formulated as:

$$\mathrm{ARN} = \frac{\sum_{i=1}^{k} \frac{T_i}{\mathrm{TF}_i}}{K} \tag{2}$$

6 Experimental Evaluation

6.1 Experimental Set-Up

To verify the performance of the proposed approach, experiments have been performed on MATLAB R.2015B and Windows 8.1. Eight different types of noises have been used for identification. These noises are taken from Aurora2 database. Experiments have been performed in three phases. Two features LPC and MFCC are used.

In first phase, features of noise (Aurora2) are directly extracted and are used for training and testing of the classifier. 60% data are used for training, and remaining are used for testing of the classifier. Classifier is trained only in first phase, and in remaining phases, the same classifier is used for testing purpose. In second phase, selected noises are mixed with IndicTTS database at 0 dB SNR value and named as noisy speech. Using VAD G.729 method, noisy frames are extracted from noisy speech and their features are calculated. These features are used for identifying the type of noise by the classifier. In third phase, subset of noisy frames is extracted using the procedure noise_extract(S). LPC and MFCC features of these frames are calculated which are later used to identify the type of noise by classifier.

6.2 Results and Discussion

In Fig. 1, average accuracies of random forest classifier for all three phases are shown. Two points are observed with this figure:

1. Average accuracy is the highest for the first phase of the experiment. It means that noise can be better classified if the noisy frames are extracted correctly.

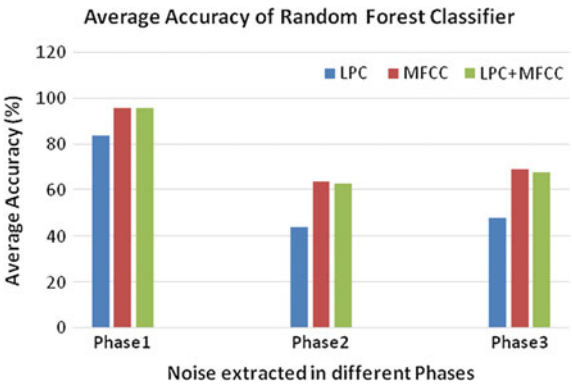


Fig. 1 Average accuracy of random forest classifier

2. Random forest classifier gives more accurate result using MFCC feature rather than using LPC alone or LPC and MFCC both. It shows that accuracy of the classifier may not be enhanced by fusing LPC and MFCC.

In Fig. 1, it is found that the performance of the same classifier is higher for the Phase3 compared to the Phase2 of the experiment. It confirms that in Phase3, noisy frames are extracted more appropriately than Phase2.

Figure 2 shows that for selected types of noises, random forest classifier performs better using MFCC feature rather than LPC feature. It is also observed that this classifier performs better for noises like airport, babble, car and station noise with MFCC feature while for others with LPC and MFCC together.

By comparing Figs. 2 and 3, it is observed that the classification accuracy of random forest classifier is better for noises extracted using procedure noise_extract (S) rather than using VAD G.729.

In Fig. 4, average accuracies of SVM classifier for all eight types of noises are shown. Average accuracy is the highest for first phase of the experiment. It means that noise can be better classified if the noisy frames are extracted correctly.

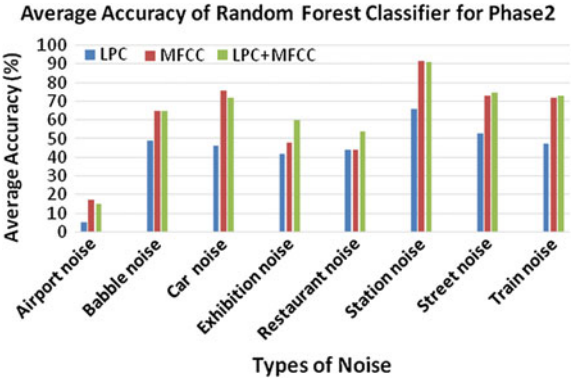


Fig. 2 Average classification accuracy for selected types of noise using random forest in Phase2

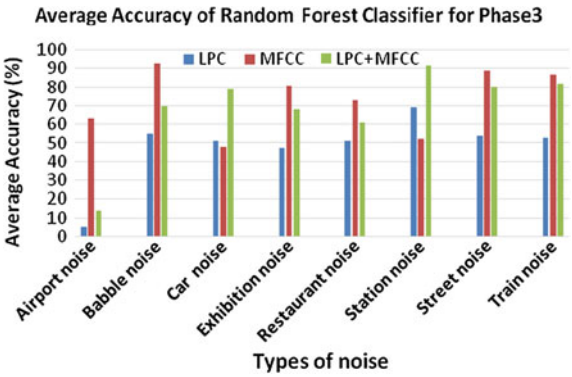


Fig. 3 Average classification accuracy for selected types of noise using random forest in Phase3

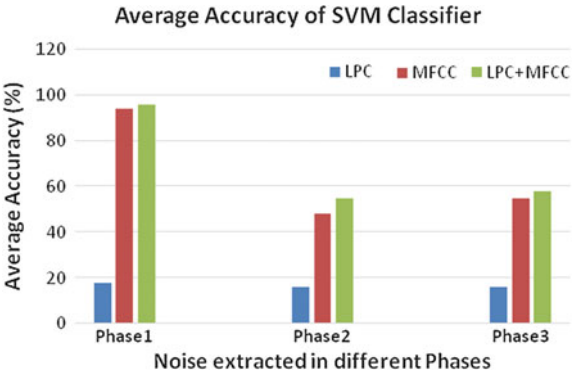


Fig. 4 Average accuracy of SVM classifier

By comparing Figs. 2 and 5, it is observed that the performance of random forest classifier is better than the SVM classifier for Phase2 of the experiment.

By comparing Figs. 3 and 6, it is observed that the performance of random forest classifier is better than SVM classifier for Phase3 of the experiment.

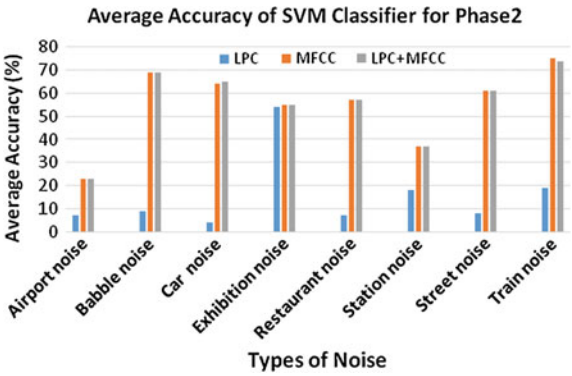


Fig. 5 Average classification accuracy for selected types of noise using SVM classifier in Phase2

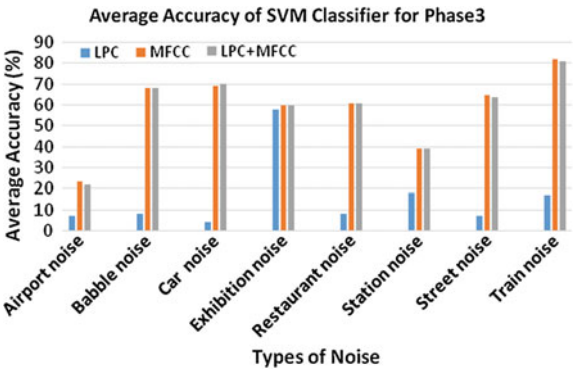


Fig. 6 Average classification accuracy for selected types of noise using SVM classifier in Phase3

Average accuracy for random forest classifier in Phase1 of the experiment is 96%. It is obtained using MFCC feature alone/LPC and MFCC features together. Average accuracies for random forest classifier are 64 and 69% for Phase2 and Phase3, respectively. Maximum accuracy is obtained using MFCC feature. Thus, average performance of the random forest classifier increases 5% from Phase2 to Phase3, respectively. It confirms that noise extracted using procedure noise_extract (S) is more accurate than VAD G.729 procedure.

Average accuracy for SVM classifier in Phase1 of the experiment is 96% which is obtained using LPC and MFCC features together. Average accuracies for SVM classifier are 55 and 58% for Phase2 and Phase3, respectively. Maximum accuracy is obtained using LPC and MFCC feature together.

Thus, average performance of the SVM classifier increased 3% from Phase2 to Phase3. It again confirms that noise extracted using procedure noise_extract(S) is more accurate than VAD G.729 procedure.

The maximum average accuracy of random forest classifier is 9 and 11% higher than SVM classifier for Phase2 and Phase3, respectively. Thus, one can conclude from the experiment that as noise in speech can better be identified through selection of appropriate noisy frames as shown in Table 1.

Table 1 Classification results of random forest and SVM classifiers for different phases of noise

Classifier	Random forest						SVM					
Features	LPC		MFCC		LPC + MFCC		LPC		MFCC		LPC + MFCC	
Noise	Parameters											
	RRN	ARN	RRN	ARN	RRN	ARN	RRN	ARN	RRN	ARN	RRN	ARN
Phase1	0.94	0.82	0.99	0.96	0.99	0.95	0.18	0.17	0.99	0.94	0.99	0.96
Phase2	0.54	0.43	0.75	0.63	0.75	0.63	0.15	0.16	0.60	0.55	0.61	0.55
Phase3	0.57	0.48	0.78	0.69	0.78	0.68	0.15	0.16	0.61	0.58	0.62	0.58

7 Conclusion

This paper presents a new framework for noise identification in speech. As noise in speech can better be identified through appropriate noisy; therefore, in this paper, an approach for extracting appropriate set of noisy frame is also proposed. Random forest classifier performs best with MFCC feature alone while to give best performance SVM needs both LPC and MFCC together. The overall performance of random forest is better than SVM for noise identification. Thus, random forest is better choice for noise identification with respect to accuracy as well as computational cost.

References

1. Selina Chu, Shrikanth Narayanany and C.-C. Jay Kuo, "Composite-DBN for Recognition of Environmental Contexts," Proc APSIPA Hollywood CA, 2012.
2. Y. Li, "A classification method for environmental audio data," 2nd International Conference on Advanced Computer Control (ICACC), pp. 355–361, 2010.
3. Y. Toyoda, J. Huang, S. Ding, Y. Liu, "Environmental sound recognition by multilayered neural networks," Proceedings of the Fourth International Conference on Computer and Information Technology, CIT'04, IEEE Computer Society, Washington, DC, USA, pp. 123–127, 2004.
4. Pradeep K. Atrey, Namunu C. Maddage and Mohan S. Kankanhalli, "AUDIO BASED EVENT DETECTION FOR MULTIMEDIA SURVEILLANCE," Acoustics, Speech and Signal Processing (ICASSP), pp. 813–816, 2006.
5. Hector Lozano, Inmaculada Hernaez, Artzai Picon, Javier Camarena and Eva Navas," Audio Classification Techniques in Home Environments for Elderly/Dependant People", International Conference on Computers for Handicapped Persons (ICCHP), pp. 320–323, 2010.
6. Byeong-jun Han and Eenjun Hwang, "IEEE International Conference on Multimedia and Expo (ICME)," pp. 542–545, 2009.
7. D.D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401 no. 6755, pp. 788–791, 1999.
8. Christian Kraetzer, Andrea Oermann, Jana Dittmann and Andreas Lang, "Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification," Proceedings of the 9th workshop on Multimedia & security, pp. 63–74, 2007.
9. C. Kraetzer and J. Dittmann., "Mel-cepstrum based steganalysis for voip-steganography," E. J. Delp and P. W. Wong, editors, Security, Steganography, and Watermarking of Multimedia Contents IX, Electronic Imaging Science and Technology, SPIE Vol. 6505, San Jose, CA, USA, SPIE and IS&T, SPIE, 2007.
10. Ho, Tin Kam, "Random Decision Forests (PDF)," Proceedings of the 3rd International Conference on Document Analysis and Recognition, pp. 278–282, 1995.
11. Ho, Tin Kam, "The Random Subspace Method for Constructing Decision Forests", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 832–844, 1998.
12. R. Shapire, Y. Freund, P. Bartlett, and W. Lee. "Boosting the margin: A new explanation for the effectiveness of voting methods," Annals of Statistics, pp. 1651–1686, 1998.
13. Leo Breiman, "Bagging predictors", Machine Learning, pp. 123–140, 1996.