IDENTIFICATION OF RECORDED AUDIO LOCATION USING ACOUSTIC ENVIRONMENT CLASSIFICATION

[1]Tejashri Pathak,[2] Devidas Dighe.

[1]*Student, Electronics & Telecommunication Department, Matoshri College of Engineering and Research Centre, Nashik, Maharashtra, India*

[2]*Professor, Electronics & Telecommunication Department, Matoshri College of Engineering and Research Centre, Nashik, Maharashtra, India*

ABSTRACT

There are many artifact and different distortions present in the recording. The reverberation depends on the volume properties of the room and it causes the calumniation of the recording. The background noise depends on the unnecessary audio source activities present in the evident recording. For audio to be considered as proof in a court, its authenticity must be verified. A blind deconvolution method based on FIR filtering and overlap add method is used to estimate reverberation time. Particle filtering is used to estimate the background noise. Feature extraction is done by using MFCC approach. The 128 Dimensional feature vector is the addition of features from acoustic reverberation and background noise and the higher order statistics. SVM classifier is used for classification of the environments. The performance of the system is checked using audio recordings dataset. The SVM classifier provides best results for the trained dataset and moderate results for untrained dataset.

Keyword: - *Reverberation, Background noise, Spectral Subtraction, Particle Filter, MFCC, SVM.*

1. INTRODUCTION

Audio forensics is the field of forensic science relating to the acquisition, analysis, and evaluation of sound recordings that may finally be presented as tolerable evidence in a court of law or some other official venue. Audio forensic investigations are of three types, authentication, enhancement and interpretation. Authentication stage verifies the originality by physical examination. Enhancement stage is used for noise reduction of the recorded audio, to increase the understanding. In the interpretation, the environment and the background noises present in the recording are estimated.

The usual steps for forensic audio examination are [1] 1. Physical Inspection: checks the condition and properties of the audio recording medium. 2. Critical Listening: Listen the entire recording and estimates the editing with the recording. 3. Spectrogram is used to identify the editing in the original recording. Digital audio voice recorders are frequently used to record and produce audio for use as evidence. Often times the audio signal can be poor quality causing precision issues within the recorded dialog. The question of authenticity becomes more complicated with digital recordings because the evidence of tampering or alterations is most difficult to discover. There are a number of methods such as Electrical Network Frequency (ENF) analysis, Time and frequency domain analysis, Pattern Recognition Systems and model driven approaches to calculate reverberation time studied in past.

The availability of powerful, sophisticated, and easy-to-use digital media manipulation tools has made authenticating the integrity of digital media even more difficult. Audio forensics has traditionally focused on analog magnetic tape recordings by relying on analog recorder fingerprints, such as head switching transients, mechanical splices, and overdubbing signatures, to determine the integrity of the recording. The question of authenticity becomes more complicated and challenging for digital recordings because digital recorders do not leave such traces in the recording.

This paper presents a mathematical modeling of audio recording to estimate the acoustic location of the original recording. The acoustic environment signature consists of features generated by acoustic reverberation and background noise. This 128 D feature vector helps us to estimate the location of the recording.

The rest of the paper is organized as follows; Section II explains the methodology used to estimate the acoustic location. Section III deals with the experiments conducted and performance analysis of the system under trained and untrained datasets. Section IV discusses the results of the whole system. Finally, section V provides conclusion and future areas to do research.

2. METHODOLOGY

Here a statistical technique based on spectral subtraction is used to estimate the amount of reverberation and nonlinear filtering based on particle filtering to estimate background noise. A blind dereverberation method based on spectral subtraction and inverse filtering is used to estimate the reverberation component. Both the background noise and the reverberation components are used for feature extraction. A 128-D feature vector consisting of MFCC, LMSC, and higher order statistics is used to characterize the acoustic environment. The SVM based classifier is used for the AEI.

The input audio recording is selected from all recordings in a database. The database consists of recordings of speeches in different environments. The recording location is considered to be unknown.

The proposed system consists of four steps

1. To calculate Acoustic Reverberation

2. To measure Background noise.

3. Combined Feature Extraction.

4. Classification

The background noise and the reverberant components anticipated from the test-recording are merged to obtain a feature vector characterizing the acoustic environment.

2.1 Input Audio and Preprocessing

Input is a unenhanced speech file in .wav (Waveform) format. In this case rounding is done towards positive infinity. The impulse response dataset of all locations is added from AIR dataset. Now take a speech file whose location is to be evaluated and sample the signal in discrete format. The sampling frequency should be 16 kHz as the human audible range is between 2 kHz to 20 kHz.

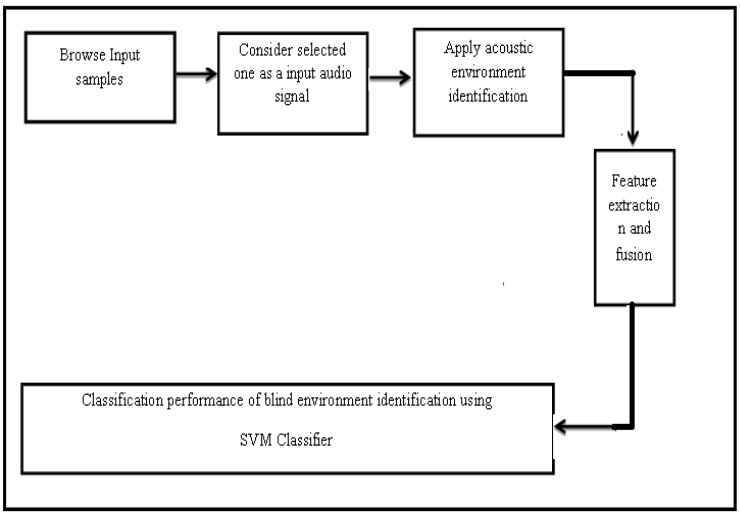

Fig -1: Block diagram of proposed system

2.2 Acoustic Reverberation Estimation

Acoustic reverberation is calculated using blind dereverberation algorithm. Speech signals recorded in hands-free situations are corrupted by room reverberation. Dereverberation is the method for automatically cancelling the reverberation effect to recover the quality of speech.

2.2.1 Generation of Blind Deconvolution Speech

The impulse response consists of three components direct signal, early reflections and late reflections. The dry signal is the first part in the original signal. The reverberation is estimated by using FFT filtering method. S_{rev} is the convolution of dry signal s(t) and the impulse response of the location. FFT based filtering is done using overlap-add method. This method works only for FIR filters.

The overlap add method is expressed in frequency domain by the equation

$$\boldsymbol{Y(z) = \left(h(1) + h(2)z^{-1} + \cdots + h(nh+1)z^{nh}\right)S(z)} \quad 2.1$$

In signal processing overlap add method is an efficient way to estimate the discrete convolution of very long signal x[n] with FIR filter h[n].

The first step consists of dividing the reverberation signal into blocks of particular length. The length of each block is 20 ms. the block length is then rounded to nearest integer value. In second step overlapping is done. The standard overlapping is 50%. The overlapping is done as linear convolution is always longer than the original sequence.

2.2.2 RT Calculation based on Room Impulse Response by Schroeder Method

Here a single measurement yields a decay curve that is identical to average over infinitely many decay curves. The range of RT estimation using this method is 0.2s to 1.2s.

The sound Decay Rate denoted by D_R and expressed in decibel (dB).

$$\boldsymbol{D_R = 60\ dB / RT_{60}} \quad 2.2$$

This method provides RT estimation over time for input x using framewise processing. The equation consists of reverberation signal, block size of overall processing scheme and overlap function gives the frame shift at each

interval. The region over which RT is calculated is [-5,-35] in dB. To extract the reverberation from the curve linear least square fitting method is used.

Estimated RT is frame shift divided by sampling frequency over the particular time interval. Blind deconvolution is expressed in terms of energy decay curve.

The 60D feature vector is generated at this stage. This feature vector consists of 30D features from Mel Frequency Cepstral coefficients (MFCC) and 30D features from Logarithmic Mel Cepstral Coefficient (LMSC).

The reverberation is estimated by using maximum likelihood estimator [12].

$$L(y;a,\sigma) = \left(\frac{1}{2\pi a^{(N-1)}\sigma^2}\right)^{N/2} \times exp\left(\frac{-\sum_{n=0}^{N-1} a^{-2n} y(n)^2}{2\sigma^2}\right) \quad 2.3$$

Next is to take logarithm of this equation followed by differentiation.

The score function is S_a

$$S_a = -\frac{N(N-1)}{2a} + \frac{1}{a\sigma^2}\sum_{n=0}^{N-1} n a^{-2n} y(n)^2 \quad 2.4$$

When the score function is equated to zero we get the maximum likelihood equation.

2.3 Spectral Estimation

Spectral estimation, in statistics and signal processing is an algorithm that estimates the strength of different frequency components (the power spectrum) of a time-domain signal. This may also be called frequency domain analysis. Here the amplitude and phase spectrum are extracted from the signal x in time domain. The steps for spectral estimation are 1. Calculate N point FFT of a signal. 2. Windowing: for this a standard hanning window is applied. The hanning periodic window returns n point symmetric hanning window in a column vector. 3. Calculate FFT of a windowed signal. 4. Calculate the number of unique points. 5. Since the FFT is symmetric over the interval so there is no need of second half. 6. For amplitude spectrum take the magnitude of FFT (x). 7. DC and Nyquist Correction: As FFT is symmetric instead of calculating the magnitude for entire vector so they calculate it for result and then double it. But DC and Nyquist components are not represented twice and therefore we divide them by 2 for compensation. 8. Convert magnitude spectrum to dB. 9. Calculate the phase spectrum by using four quadrant inverse tangent i.e. arctan function. 10. Then convert phase spectrum to degrees

The spectral estimation is achieved by dividing the input audio into 20 ms overlap blocks with 50% overlapping followed by time domain smoothing using 2048-point Hamming window, a frequency domain transformation using DFT, filtering using Mel-filter bank and rescaling using the logarithmic function.

2.4 Background Noise Estimation

The background noise was reduced by using noise cancelling microphones. As the real world noise is dynamic and nonlinear in nature so particle filtering approach is invented to solve this problem. Particle filters (PF) or sequential Monte Carlo methods, initially developed for typical tracking applications like pursuing airplanes in radars. They are used for the enhancement of speech features corrupted by noise. The two critical aspects in PF design are the choice of the importance or proposal density and the particle weight calculation. For calculation of weights in particle filter sequential importance sampling (SIS) type particle filter is used. The steps for sequential importance sampling are initialization, prediction and updating. The disadvantage of SIS PF is degeneracy problem where a sampling step selects a reasonable number of samples with insignificant weights. To solve this problem residual resampling is done which discards samples with small weights and maintain a constant number of samples.

2.5 Feature Extraction Stage

Feature extraction is done using Mel Frequency Cepstral Coefficients (MFCC) approach. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on

a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The mel frequency scale is linear frequency spacing below 1kHz and logarithmic spacing above 1kHz.The difference between the Cepstrum and the mel-frequency Cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal Cepstrum. The motivation behind selecting MFCC for feature extraction is that cochlea of human ear is receptive to only mel scale frequencies.

The complete 128 D feature vector consists of 60D features from acoustic reverberation step, 60D features from background noise step and 8 higher order statistics such as mean, variance, skew, and kurtosis. The 60D acoustic reverberation features consists of 30D features from MFCC and 30D features from LMSC (Logarithmic Mel Cepstral Coefficients). The 60D feature vector of background noise step consists of 30D LMSC and 30D MFCC.

2.6 Classification

The classification step consists of identifying the locations of the recordings. The trained 128 D feature vector is provided as an input to SVM classifier. The classifier then detects the location where the recording was actually done. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. The SVM classifier uses a RBF (Radial Basis Function) which is also called as Gaussian Kernel as this kernel provides improved accuracy compared to linear and polynomial based classifiers.

$$K(x, x') = exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad 2.5$$

Where $\|x - x'\|^2$ = Squared Euclidean Distance.

3. EXPERIMENTS

This section explores the performance of the system for various recordings done in different locations. The assumption is that we don't know the location of the recording. This is the blind detection of the location. The performance of the system is checked for one trained and two untrained datasets. The trained dataset consists of recordings from the IEEE scene identification database of audio recording locations. The second dataset is the untrained dataset of recordings from free sound software. Third dataset consists of recordings done by myself in the field. Fig below shows the results of the proposed system for trained dataset.

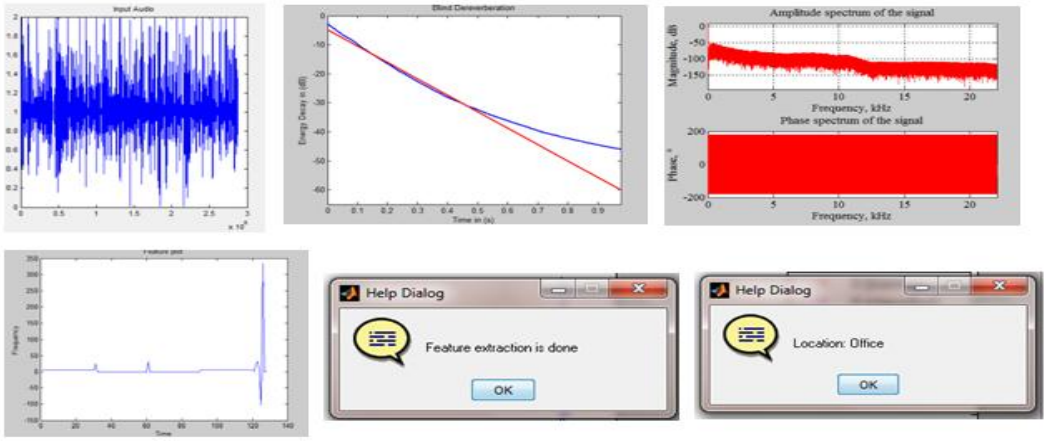

Fig -2: Results of trained dataset (a) input audio (b) blind dereverberation (c) spectral estimation (d) feature plot (e) feature extraction (f) SVM classification

Table below shows the comparison of trained and untrained datasets.

Table -1: Results of trained dataset

Location	Trained Dataset	
	Energy Decay	SVM accuracy
Bus	48 dB	80%
Office	47 dB	100%
Park	47 dB	80.23%
Restaurant	48 dB	83.6%
Tube	48 dB	58%

Table -2: Results of untrained dataset

Location	Untrained Dataset			
	Dataset 2		Dataset 3	
	Energy Decay	SVM accuracy	Energy Decay	SVM accuracy
Bus	48 dB	80%	48 dB	50%
Office	47 dB	80%	48 dB	64%
Park	48 dB	40%	48 dB	50%
Restaurant	48 dB	38.75%	48 dB	34.6%
Tube	48 dB	50.5%	48 dB	52%

4. DISCUSSIONS

The proposed system shows 100 % accuracy in the case of recording done in office. The SVM classifier gives best performance in terms of trained dataset. The reverberation does not show change in energy decay. It is almost same for every location in trained dataset.

The dataset 2 is taken from the freesound recordings software. This dataset also provides moderate results. But the performance drops in recordings at park and restaurant as some restaurants are more crowded than others. In this dataset some locations have no background noise so this is shown in feature plot by small amplitude.

The third dataset consist of recordings done at field such as office and travelling in a bus. This dataset provides moderate results in terms of location detection. The location such as restaurant is detected but with less accuracy. The classifier accuracy is below 40% in some cases so we can say that the location detected is not reliable. The recording may or may not be done at that particular location.

The background noise is prominent in some cases and that provides good performance as the feature vector is complete. But the recording in which there is no or less background noise the performance of the system degrades.

5. CONCLUSION

Reverberation and background noise are used to characterize the sound environment. Background noise is calculated using a nonstationary system and estimated using particle filtering. A blind deconvolution method based on FIR filtering and overlap add method is used to estimate reverberation time. Both the above components are used for feature extraction. This work proposed a system to identify the location of the audio recording captured in a certain environment. This is a full blind identification of the location of the recording. It is assumed that the recordings are made with unchanging set of microphone or sound since it is intricate to estimate the reverberation component from nonstationary acoustic scene. The performance provided by untrained dataset is moderate so it is concluded that the system is reliable for location identification. In this work it is assumed that the recordings are made with fixed set of microphone or sound source since it is difficult to estimate the reverberation component from nonstationary acoustic location. In real life scenario the noise is random in nature so there is a scope to work on this point. The audio recordings are taken from a database so they are not real time. So it is necessary to calculate the reverberation and background noise from a real time audio recording.

REFERENCES

[1]Tejashri Pathak and Prof. Devidas Dighe, "Digital Media Authentication Method for Acoustic Environment Detection" International Journal of Innovative Science, Engineering & Technology, vol 2, issue 4, April 15, ISSN 2348 – 7968.

[2]H. Zhao and H. Malik "Audio recording location identification using acoustic environment ignature," IEEE Trans. Inf. Forensics and Security, vol. 8, no. 11, Nov 2013.

[3]Audio Engineering Society, AES43-2000: AES standard for forensic purposes – Criteria for the authentication of analog audio tape recordings (2000).

[4]C. Grigoras, "Digital audio recording analysis: The electric network frequency (ENF) criterion," Int. J. Speech Lang. Law, vol. 12, no. 1, pp. 1350–1771, 2005.

[5]D. Rodriguez, J. Apolinario, and L. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Trans.* Inf. Forensics Security,vol 5, no. 3, pp. 534–543, Sep.2010.

[6]A.Oermann, A. Lang and J. Dittmann, "Verifier Tuple for audio forensic to determine speaker environment" in proc. ACM Multimedia and Security Workshop , New York,NY, USA, pp. 57-62, 2005.

[7]R. Malkin and A. Waibel, "Classifying user environments for mobile applications using linear autoencoding of ambient audio, " in Proc. IEEE Int. Conf. on Acoustic., Speech, and Signal Processing, vol. 5, pp. 509-512, 2005.

[8]R. Buchholz, C. Kraetzer and J. Dittmann, "Microphone Classification using Fourier Coefficients," in Lecture Notes in Comput. Sci. Berlin/Heidelberg, Germany: Springer vol 5806/2009, pp. 235-246, 2010.

[9]D. Garcia - Romero and C. Espy – Wilson, "Automatic acquisition device identification from speech recordings," J. Audio Eng. Soc., vol. 124, no. 4, pp. 2530-2530, 2009.

[10]Y. Panagakis and C. Kotropoulos, "Automatic telephone handset identification by spares representation of random spectral features," in Proc. Multimedia and Security, pp. 91-96 2012.

[11]C. Kraetzer, K. Qian, M. Schott and J. Dittmann, "A context model for microphone forensics and its application in evaluation," Proc. SPIE Media Watermarking, security, and Forensics III, vol. 7780, pp. 1-15, 2011.

[12]R. Ratnam, D. Jones, B. Wheeler, W. O'Brien, "Blind estimation of reverberation time,"J. Acoust. Soc. Amer., vol. 5, no. 114, pp. 2877-2892, 2003.

[13]G. Soulodre, "About this dereverberation business: A method for extracting reverberation from audio signals," in Proc. AES 129th convention, San Francisco, CA, USA, 2010.