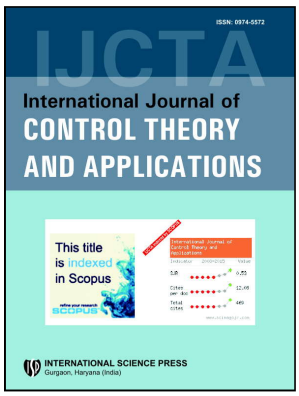


International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 14 • 2017

Clustering based Noise classification using Speech Data

Shanthi Therese S.[1] and Chelpa Lingam[2]

[1] Research Scholar , RAIT Affiliated to the University of Mumbai, Email: shanthitherese123@gmail.com
[2] IEEE Member Pillai HOC College of Engineering & Technology Rasayani, Mumbai, India Affiliated to the University of Mumbai, Email: chelpa.lingam@gmail.com

Abstract: In this work, main objective is to classify the noise present in the speech data. It is always a human nature to predict a voice associated with its background noise. Speech and Noise are integrated components. The noise part of the given signal includes background noises (car noise, kitchen noise, fan noise, street noise, etc.) and/or speech from multiple speakers. The periodicity and aperiodicity characteristics help to separate speech and the noise components. For testing purpose various environment noise are considered in this work. NOISEX database includes different kind of environment noise like babble, street, station, airport, etc. All different noises are recorded at various SNR levels like 0dB, 5dB, 10dB, 15dB. Using different classifier result analysis were obtained. Average accuracy of up to 80% is achieved in the recognition stage of different noises.

Keywords: Noise Analysis, Periodic and Aperiodicity features, DBScan

1. INTRODUCTION

Speech recognition is an important ongoing research work in the pattern recognition area. Though the challenges in speech recognition area have been attempted by many researchers, a thought for noise analysis involved in the speech is still an untouched area. In this work, we have analyzed different noises that could become a part of the speech. When we record speech noise is always an integral component of sound. Characterization of noise is always a challenging work among the researchers of Automatic Speech Recognition (ASR). To improve the robustness of ASR noise analysis plays an important role. Through the periodic and a periodic characteristics involved in the noise we can always broadly classify particular noise as periodic or Aperiodic noise. Periodic noise is the one in which pitch frequency can be obtained by analyzing the periodicity present in the noise spectrum. Speech is suppressed and Noise is enhanced in this work so that the first stage of noise classification as either periodic or aperiodic can be obtained. In the second stage , periodic and aperiodic group of noises are analyzed by using clustering algorithm. Since noice is always considered as random variations, here finding some pattern in the randomness helped us to classify eight noises present in the NOISEX database into two groups . One is periodic and the other one is aperiodic. Understanding of this category, will lead the researchers to identify the background at which the given audio is recorded. In speech recognition area, there are many feature extraction algorithm are involved to extract the features very efficiently. But in this work the nature of

pattern to be recognized is unique in each type of noise. For example train noise , we can expect periodic variations from the train. moving sound. So here we can employ pitch extraction algorithm. Car noise, we are actually finding continuous vibrations based on the vehicle moving speed and interrupting street noise like horn of other vehicles. If there is any noticeable peak points, occurred in the speech we can classify it as car noise. In railway stations and airport noise, we can expect periodic announcements in a very loud voice. These announcements are actual speech signals which are periodic. These speech signals are prominent in the railway stations. So this noise is also grouped under Periodic noise. In exhibition and babble noise category , we have used voice activity detection algorithm. If voice activity involved in the speech is more than a threshold parameter, we can label that noise as babble or exhibition noise. In restaurant noise we can always expect the presence of some music or TV background voice excitation. If this can be measured it can be categorized as restaurant noise. In absence of this there is high probability of any algorithm to misclassify babble as exhibition noise and vice versa. From the above mentioned parameters in section III we have proposed a model to classify noise. Section IV deals with the result analysis. Conclusions derived from this study are discussed in Section V.

2. RELATED WORK

W. Van summers et. al carried acoustical analyses on a set of utterances produced in a masking noise environment of 80,90 and 100 dB. Perceptual experiments were carried to measure the intelligibility of utterances in both quiet and loud noise scenario. [1] Nitish Krishnamurthy and John H.L. Hansen have developed a framework to detect underlying structure of babble noise. This work contributes much to improve the robustness of speech recognition in noise. [2] Wang W.Y in his proposed work, introduced a method to analyze the signal and identify the noise sources like colliding sources, revolving machines, moving vehicles, computer fans etc. . An invertible transform method is proposed in this work. [3] The effect of additive noise on speech amplitude , a quantitative analysis was proposed by Qifeng Zhu , et. al. to estimate speech spectra from noisy conditions. From the estimated spectra, MFCC features were extracted to form front end.[4] Chanwoo Kim and Richard M. Stern developed enhanced power normalized cepstral coefficients. feature extraction technique . In this authors have employed "asymmetric nonlinear filtering" to estimate the level of background noise for each time frame and frequency bin. Thus slowly changing frequency components are identified easily. [5] Sanjay P. Patil and John N Gowdy proposed a noise estimation technique based on spectral sparsity. Frame to Frame phase difference used as a means of detecting the noise. Noise estimation for non stationary noises like babble , restaurant and subway noise are handled using this noise estimation algorithm.[6] Bin Gao and Wai Lok Woo proposed a model called as wearable audio monitoring. In this work , different audio features like ZCR , Energy Entropy(EE) , Short Time Energy(STE), Spectral Energy(SE) were extracted and used for classifying block of sound and speech and classification of both environmental and speech sounds.[7] Jian Zhou and et. al. proposed a non negative matrix factorization algorithm to extract phonem bases from whispered speech. Noise bases were obtained from training using the conventional non-negative matrix factorization.[8] In speech recognition , feature extraction plays a vital role in improving the recognition accuracy, Though noise is an unwanted signal present along with the speech, which is categorized in this work all the feature extraction techniques used in speech enhancement are to be highlighted. Most commonly used are Linear Predictive Analysis (LPA) , Linear Predictive Cepstral Coefficients (LPCC) [9] , Perceptual Linear Predictive Coefficients (PLP)[10] , Mel-frequency Cepstral Coefficients (MFCC)[11] , Relative spectra filtering of log domain coefficients (RASTA) [12]. Noise compensation algorithms were proposed by different researchers have provided substantial improvement in accuracy for recognizing speech in the presence of quasi-stationary noise. [13]-[20] . Speech recognition under traffic noise had been demonstrated by G. S´arosi , et al.[21] The implementation of Noise classification is discussed in section III.

3. PROPOSED METHODOLOGY

The most widely used approach in noise estimation involves Voice Activity Detection (VAD) based algorithms. The VAD algorithm extracts features related to harmonics of the signal like Short Time

Energy(STE), Zero Crossing Rate(ZCR) from the input signal. Threshold based comparison is made to decide a particular frame is voiced or unvoiced. VAD algorithms generally outputs a binary decision per frame where frame length is of standard time span of 20-30ms. A frame is declared to contain voice activity (VAD=1) if the measured feature value exceeds a given threshold, otherwise it is considered to be noise(VAD=0).

In this work, two stage classification is proposed. In the first stage, we analyse the input signal for periodic features. If there is periodic signal present , it is classified as periodic noise. Perioidic noise group includes, Train Noise, Airport Noise and Station Noise. Aperiodic noise includes Car noise , Babble Noise, Exhibition Noise, Restaurant Noise and Street noise.

In the second stage classification , for periodic noise further analysis based on frequency domain and time domain analysis were carried. All the peak points were estimated using amplitude based method and pitch variations were analyzed using pitch extraction algorithm. For peak estimations we have used both the Short Time Energy (STE) and Zero crossing rare (ZCR) calculations based model.

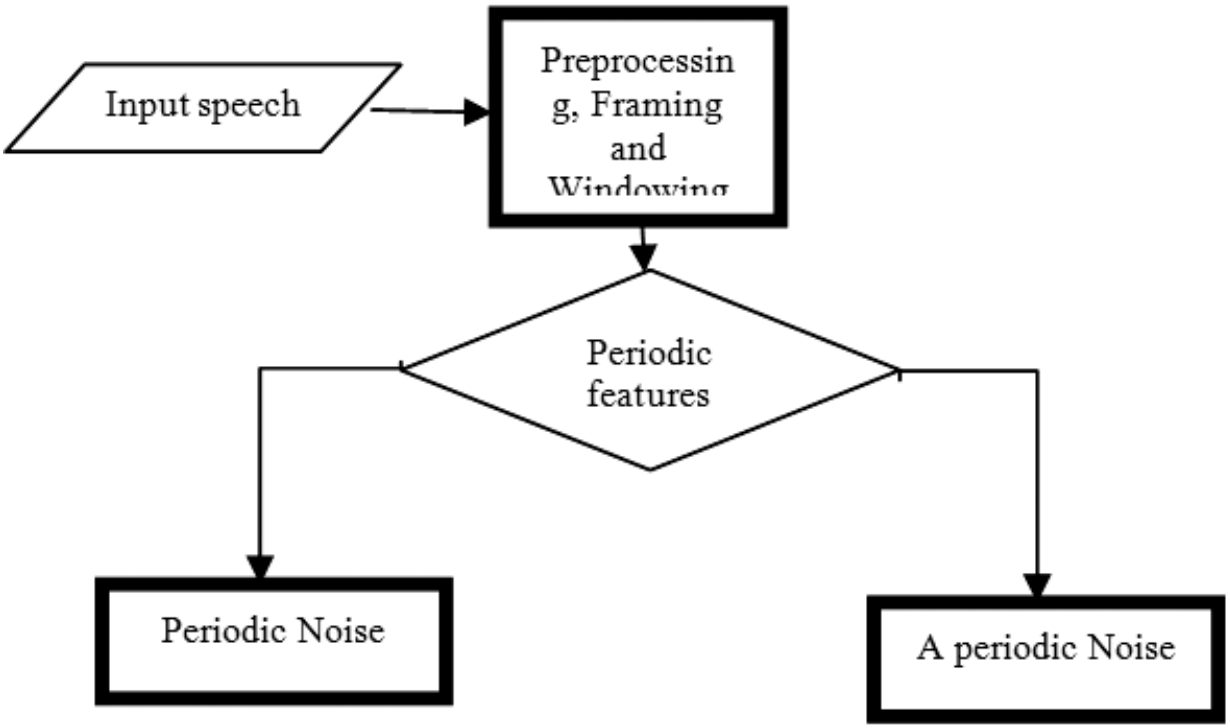


Figure 1: First Stage classification of Noise

Short Term Energy is obtained by using the framed short duration sample of 25 ms. Using this small sample of 25 ms of every segment, energy of short term can be calculated using

$$E(T) = \sum_{k=0}^{n-1} s^2(n) \tag{1}$$

Equation 1 gives the total energy present in the signal for frames 0 to n-1. Only one such frame can be modeled as extracted using a windowing function.

$$Sw(n) = S(m).W(n-m) \tag{2}$$

Equation 2 represents the speech signal duration by applying hamming window function.

The ZCR gives an indirect information about the frequency content of the signal. The ZCR high value indicates the frequency is more and if it is less the signal is changing slowly. These estimates can be used to identify the frequency variation in the signal.

$$Z(n) = 1/2N \sum_{m=0}^{n-1} s(m).w(n-m) \tag{3}$$

In equation 3 Since 2 zero crossings will be there in each cycle the denominator is 2.

In the second group of noise, to estimate the aperiodic features and calculate the exact parameter present in the signal we have used stage2 classification.

Steps involved in stage 2 classification are given below.

Input audio signal is preprocessed using neighborhood samples filtering approach. Each sample is replaced by a mean of 2 samples on the left side and 2 samples on the right side. Initially, the given speech signal is given to the input of signal filtering technique.

$$S_f(x) = \frac{\sum_{i=-2}^{2} s(x+i)}{Wi} \tag{4}$$

Where, *S(x)* is the extracted audio signal, *(S (x + i))* represents the previous and present time sample data,

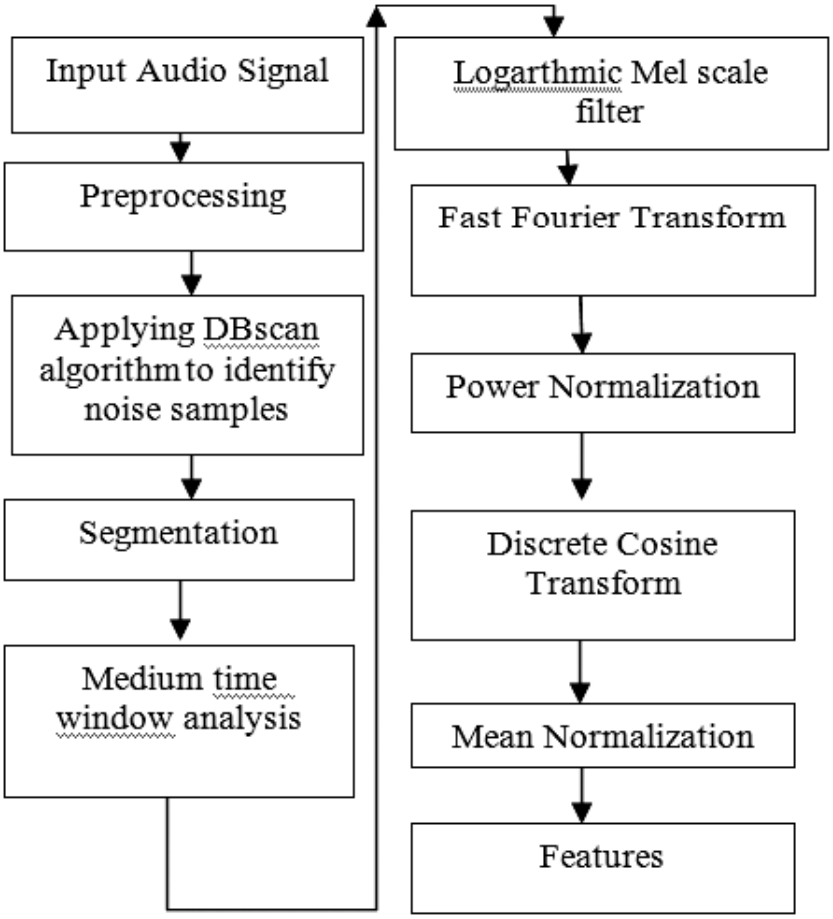


Figure 2: Second Stage classification of Noise

which will get the average value of its amplitude with respect to the filter window (w_f) and $(S_f(x))$ defines the normalized filtered audio signal output. Normalized output is scanned for noise. In this noise are outliers. All the outliers are marked by using DBScan algorithm. By using these we can group three clusters. Cluster with only voice , only noise and voice and noise overlapped. All samples clustered together were analyzed further for finding the characteristics of the noise present. These points are segregated and used for classification. Using DBscan algorithm prominent outliers can be detected exactly. These indices are marked as outlier points. All these outlier points are segregated into segmented signal. These components are actually the noise present in the signal. All these locations are collected and divided into 20 frames for short time analysis. In this short time analysis we could point easily any intensity variations and pitch variations. Using this property of clustered noise points we can classify any particular group1 and group2 noise. Features like ZCR, STE combined with the above features Data mining standard tool Weka is applied with different training set and test set. Results obtained are demonstrated in section IV.

4. EXPERIMENTS & RESULTS ANALYSIS

The noise recognition experiments were performed on wav files from the NOISEX corpus. Total training set comprised of the above specified eight categories of noise files. Each type of noise with nominal SNRs of 5dB, 10dB and 15dB were used. Total 960 Noise files of different classes were used as shown in Table 1. In that few files which includes all irrelevant features were removed. Remaining 863 files were used in training case. Since the noise can remain stationary for more than a vocal cord excitation, the analysis of spectrum and cepstrum can be carried on medium time to long time analysis window. This will reduce the computational complexity. For each noise type its unique envelope can be derived and can be compared with the test noise input. The features of Noise and Speech were stored in ARFF (Attribute Relation File Format) form. The following classification results were obtained using Data Mining tool Weka.

Table 1
Dataset
Noise at different SNR levels.

Noise	*No of Training files*	*Type*	*0dB*	*5dB*	*10dB*	*15dB*
Airport Noise	120	1	30	30	30	30
Babble Noise	120	2	30	30	30	30
Car Noise	120	3	30	30	30	30
Exhibition Noise	120	4	30	30	30	30
Restaurant Noise	120	5	30	30	30	30
Station Noise	120	6	30	30	30	30
Street Noise	120	7	30	30	30	30
Train Noise	120	8	30	30	30	30

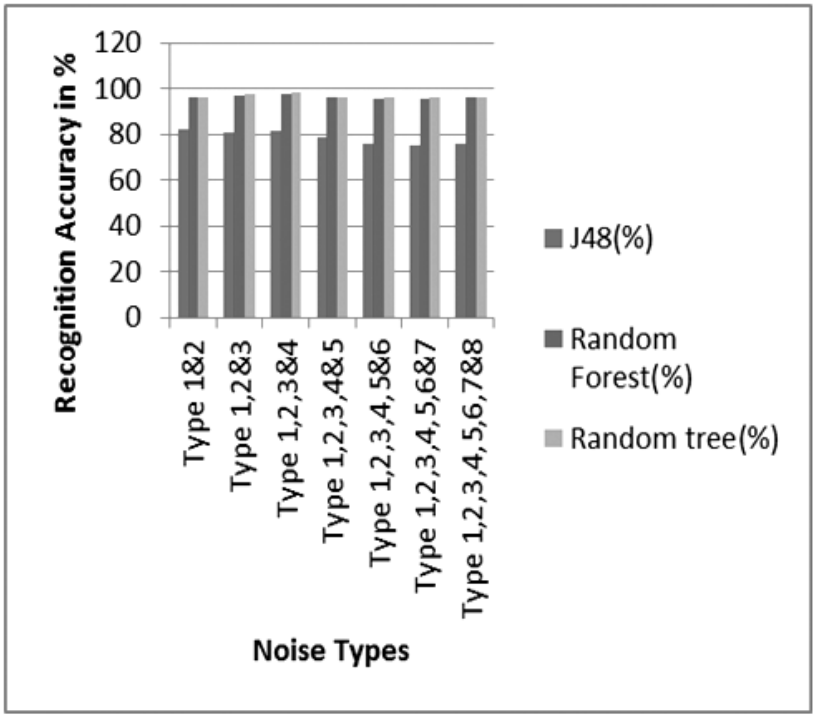
All these different classifier algorithms calculate the recognition accuracy based on True positive and True Negative assignments. TP indicates True positive, TN indicates True Negative , FN indicates False Negative and FP indicates False positive.

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{5}$$

$$= \frac{Number\ of\ true\ correct\ assessment}{Number\ of\ all\ assessment} \tag{6}$$

Table 2
Detailed Noise classification Accuracy using different classification models

Number Noise Types	*No of of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random tree(%)*
2	120	1 and 2	82	96	96.49
3	331	1 to 3	80.66	97.28	97.58
4	439	1 to 4	81.54	97.9	98.17
5	555	1 to 5	78.91	96.39	96.51
6	675	1 to 6	75.55	95.85	96
7	786	1 to 7	75.45	95.8	95.92
8	863	1 to 8	75.55	95.94	96.29



The above training file considering all the samples into the given 8 classes, the following results were obtained using the decision tree based classification algorithm J48 , Random forest and Random Tree.

Table 3
Case 1 - Recognition accuracy considering all the Eight class of noise (Only Training set)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
1-8	863	1 - 8	84.93	96.87	96.98

Case 2: Recognition accuracy considering all the Eight class of noise (at all different db levels with different labels)

a. Training set : 590 Training Features (32 classes of noise . That is 8 noises of 4 different DBs)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
32 Labels	590	1- 32	73.89	95.76	95.93

b. Test set : 197 Training Features (32 classes of noise . That is 8 noises of 4 different DBs)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
32 Labels	197	1- 32	63%	75.5%	79.5%

Case 3- Recognition accuracy considering all the Eight class of noise

a. Training set : 590 Training Features (8 classes of noise .)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
8Labels	590	1 - 8	84.74	96.44	96.61

b. Test set : 197 Test Features (8 classes of noise .)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
8 Labels	197	1- 32	74%	79%	81%

Case 4 - Recognition accuracy considering only first 2 classes of noise (restaurant , babble)

a. Training set : 170 Training Features (2 classes of noise .)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
2 Labels	170	1-2	94.74	98.23	98.23

b. Test set : 58 Test Features (2 classes of noise .)

Noise	*No of Training files*	*Type*	*J48(%)*	*Random Forest(%)*	*Random Tree(%)*
2 Labels	58	1 - 2	88%	85%	90.34%

5. CONCLUSIONS AND FUTURE WORK

In this proposed work, different noises included in the NOISEX database were analyzed. Time domain and frequency domain characteristics of each noise were used in categorizing the given noises In this work , clustering algorithm groups the samples into three clusters. Samples with only Voice, samples with only noise and samples overlapped with noise and voice. These clusters were analyzed further for the frequency variation pattern and the prominent difference were captured as features for classification. In Future work this work can be carried forward to include different type of noises. Characterization of each noise will be always encouraged to make ASR system more robust in the research area of ASR.

REFERENCES

[1] W Van summers, David B.Piosoni, Robert H.Bernacki, Robert I.Pedlow and Michael A Stokes," Effects of noise on speech production: Acoustic and Perceptual analyses", J*ournal of Acoustic Society America*, Sep 1988 pp 917-928

[2] Nitish Krishnamurthy and John H. L. Hansen, "Babble Noise: Modeling, Analysis, and Applications" ,*IEEE Transactions on Audio, Speech and Language Processing*, Vol. 17, No. 7 September 2009

[3] Wang W.Y., "New method to Analysis Noise in Speech Signal", Pacific–Asia Conference on Circuits, Communications and Systems, 2009, PACCS'09.

[4] Qifeng Zhu and Abeer Atlwan, "The Effect of Additive noise on Speech Ampllitude Spectra: A Quantitative Analysis",*IEEE Signal Processing Letters*, Vol. 9 , No.9 , September 2002 pp 275-277

[5] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4101-4104.

[6] Sanjay P.Patil and John N. Gowdy, "Use of baseband phase structure to improve the performance of current speech enhancement algorithms" , Speech Communication, Elsevier Nov 2014 ,pp 78-91

[7] B. Gao and W. L. Woo, "Wearable Audio Monitoring:Content-Based Processing Methodology and Implementation,"*IEEE Transactions on Human-Machine Systems,* vol. 44, pp. 222-233, 2014.

[8] J. Zhou*, et al.*, "Unsupervised learning of phonemes of whispered speech in a noisy environment based on convolutive non-negative matrix factorization," *Information Sciences,* vol. 257, pp. 115-126, 2014.

[9] Harshita Gupta and Divya Gupta," LPC and LPCC method of feature extraction in speech recognition system" , Cloud System and Big Data Engineering , Confluence 2016

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spokenentences," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980

[11] H. Hermansky, "Perceptual linear prediction analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[12] A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," in*Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (Albuquerque, NM)*, vol. 2, Apr. 1990, pp. 849–852.

[13] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, May. 1996, pp. 733–736.

[14] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," in*IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, May 2006, pp. 773–776.

[15] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Apr. 1997, pp. 33–42.

[16] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. M. Davis, Ed. CRC Press, 2002, pp. 219–244.

[17] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms* European Telecommunications Standards Institute ES 202 050, Rev. 1.1.5, Jan. 2007.

[18] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Nov. 2001, pp. 21–24.

[19] H. Mirsa, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *IEEE Int. Conf. Acoust. Speech, and Signal Processing*, May 2004, pp. 193–196.

[20] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in*IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 2, Apr. 1997, pp. 851–854.

[21] G. Sarosi, M. Mozs´ary, P. Mihajlik, and T. Fegyo, "Comparison of feature extraction methods for speech recognition in noise-free and intraffic noise environment," in *Speech Technology and Human-Computer Dialogue (SpeD)*, May 2011, pp. 1–8.