Investigation on the Effect of the Input Features in the Noise Level Classification of Noisy Speech

T K Dash¹²* and S S Solanki¹

¹Electronics and Communication Engineering, Birla Institute of Technology, Mesra, India
²Electronics and Telecom Engineering, C V Raman College of Engineering, Bhubaneswar, India

Received 4 September 2018; revised 17 July 2019; accepted 22 October 2019

Noise Level Estimation plays a crucial role in Speech Enhancement (SE) Algorithms. Recently, few noise estimation (NE) algorithms are developed for SE using the minimal-tracking method, but there is little research done in the noise level classification (NLC). Therefore, there is a need to identify appropriate audio features that are required for the NLC. In this paper, this problem has been addressed and seventeen audio features of the noisy speech are examined for NLC using four different types of standard and efficient classifiers such as K-Nearest Neighbor (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) classifiers. The features are first optimized to achieve the best classification performance using the Principal Component Analysis (PCA) and the Neighbourhood Component Feature Selection (NCFS) method. Finally, a comparative performance analysis is carried out by taking six different categories of real-life noisy speech signals from the standard speech database and then the best set of features are reported and the best performing classifier for the NLC is identified.

Keywords: Noise Level Estimation, PCA, Neighbourhood Component Feature Selection, Speech Enhancement, Noise Level classification

Introduction

The speech processing algorithms have a wide range of applications in the field of SE¹², speech recognition³, emotion recognition⁴, and speech communications. In these applications, the NE and NLC are performed for the reduction of background noise and for the improvement of quality, and intelligibility of noisy speech signal. The SE schemes require the knowledge of the level of noise in the noisy speech signal. Hence, there is a requirement of proper classification of noisy speech. Depending upon the level of noise in speech, the enhancement technique can be carried out accordingly. Therefore, NE plays an important role in the overall performance analysis. The review of related literature in the area of NE reveals that the research work related to the classification of background noise level is few⁵. Recently, there is reported work in the literature on the NLC of babble noise⁶ in which the proper selection of audio feature plays an important role in it. When the only single-channel noisy signal is available without any reference of a clean speech signal, the analysis of audio features becomes very crucial for the estimation of the noise level. The audio features are used to represent the values of audio data samples with fewer values by removing irrelevant information. These features can be broadly classified into different categories based on the statistical properties, spectral shape, technical/signal properties, intensity properties⁸. This paper presents a thorough analysis of the several audio features and their relationship to changes in noise level. The organization of the paper proceeds as follows the materials and methods relating to the speech data set, audio features selection, PCA and NCFS methods are presented in Section 2. In Section 3, the methodology of implementing different types of classifiers and the evaluation process is dealt with. The simulation results and the comparative performance analysis are discussed in Section 4. Finally, the conclusion and scope for further research work are provided in Section 5.

Materials and methods

In this section, the details of the speech data set and the classification methods are discussed.

Speech Dataset

The NOIZEUS speech data set⁷ is used for the
training and testing of the classifiers. It is a noisy speech database that contains 30 IEEE sentences recorded in the voice of three male and three female speakers and then mixed with the eight different real-world noises (suburban train noise, babble, car, exhibition hall, restaurant, street, airport, and train-station noise) at different SNR levels. It contains a total of 960 noisy speech files and 30 clean files recorded at 8 kHz sampling frequency.

Audio Features
For classification of noisy speech, several audio features are used and the details are discussed in this section.

Spectral Centroid ($f_1$)
It provides information about the center of gravity of spectral energy and expressed in as the ratio between the frequency-weighted sum of the power spectrum and its unweighted sum.

Spectral Crest Factor ($f_2$)
It is a measure of the tonalness of sound and is written as the ratio between the highest value of the magnitude spectrum and the addition of its magnitude spectrum.

Spectral Decrease ($f_3$)
It gives information about the steepness of the decrease of the spectral envelope with respect to the frequency range.

Spectral Flatness ($f_4$)
It is expressed as the ratio between the geometric mean and arithmetic mean of the magnitude spectrum.

Spectral Kurtosis ($f_5$)
It determines whether the spectral magnitude distribution is spread like a gaussian distribution or not.

Spectral MFCC ($f_6$)
The Mel Frequency Cepstral Coefficients (MFCC) provide an overall description of the structure of the spectral envelope. These are computed by using the mel warped spectrum with a bank of overlapping band-pass filters, and then applying the logarithm and calculating the Discrete Cosine Transform on each of the frequency resulting bands.

Spectral Pitch Chroma ($f_7$)
It is related to the expression of the pitch perception as the frequency ratio using fundamental frequency. It is calculated by grouping the audio signal block into semi-tone bands and a measure of salience is computed in each band, and then the sum across all bands with respect to a specific pitch class is calculated.

Spectral Rolloff ($f_8$)
It is the measure of the bandwidth and is defined as the frequency bin below which the accumulated magnitudes of the frequency spectrum will reach a specific percentage (85 %) of the overall sum of magnitudes.

Spectral Skewness ($f_9$)
The spectral skewness is the assessment of the symmetry of the distribution of the spectral magnitude values over their arithmetic mean.

Spectral Slope ($f_{10}$)
It is an estimation of the slope of the spectral shape.

Spectral Spread ($f_{11}$)
It is specified as the instantaneous bandwidth and gives the information regarding the concentration of the power spectrum around the spectral centroid.

Spectral Tonal Power Ratio ($f_{12}$)
The tonalness of a signal means the amount of tonal components present in the signal with respect to the noisy components. Tonalness is a measure related to sound quality and inversely proportional to the amount of noisy components. It is calculated as the ratio between the tonal power and the overall power.

Teager Energy ($f_{13}$)
Teager Energy (TE) gives the proper modeling of airflow properties of vocal tract and glottis. The expression of the TE calculation of a speech signal $y(n)$ is given in Equation (1).

$$f_{13}[y(n)] = y^2(n) - y(n+1)y(n-1) \quad \ldots (1)$$

Entropy ($f_{14}$)
It is related to the probability of the samples in a speech signal. The absolute value of entropy is expressed in Equation (2), where $p(n)$ is the probability of occurrence of the $n$th sample value.

$$f_{14}(n) = \Sigma_{n=1}^{N} p(n) \log_10 (p(n)) \quad \ldots (2)$$

Zero Crossing Rate ($f_{15}$)
It indicates the number of reversals in the sign of
the magnitude of $y(n)$ inconsecutive audio samples.

**Mean ($f_{16}$) and Standard Deviation ($f_{17}$)**

Mean is the average amplitude value of speech samples in the time domain. Standard deviation measures the spreading of the input signal from the mean value.

**Principal Component Analysis**

The PCA is used to maximize the variance in the features of the data set and to generate a new set of uncorrelated features to increase the classification accuracy\textsuperscript{11,12}. Normally, the PCA is performed using following steps: (i) data normalization, (ii) removal of the zero-mean data, (iii) projection of the data into the direction of a unit vector to get the maximum variance.

**Neighbourhood Component Feature Selection**

The NCFS is the nearest neighbor-based feature weighting algorithm, which is based on the maximization of the expected leave-one-out classification accuracy\textsuperscript{13}. It gives information about the features which are relevant in the classification task. The performance of NCFS is indifferent to the kernel width and the regularization parameter. The fscnca method mentioned in Matlab Machine Learning Toolbox is used in the proposed algorithm for the NCFS implementation.

**Results and Discussion**

**Proposed Evaluation Strategy**

The steps for evaluation of the classification algorithms is shown in Figure 1. For simulation the six types of the real world noises like babble, car, exhibition hall, street, airport, train-station noises are taken from the NOIZEUS database. In total, seventeen features are extracted from the 720 noisy speech files. After feature extraction, the feature values are normalized by dividing each feature value by its corresponding maximum value of that group. The standard length of the noisy speech files is of 3 seconds duration. The mean and absolute values of all the features are used in the calculation.

**Evaluation Measures**

Classification Accuracy (CA), F-measure (Fm), Prediction Speed and Training Time are used for the performance evaluation of classification.

**Classification Accuracy**

The CA measure is to identify the correct assignment of noisy speech signals into the corresponding noise subcategory of noise levels\textsuperscript{14}. F-measure

$Fm$ is used to evaluate the number of correct signals matched from class $j$ in cluster $c$. It is calculated by first finding the values of precision ($Pr$) and recall ($Rc$) measures as expressed in Equation (3), where $m_{j,c}$ is the number of the class member $j$, which is present in cluster $c$, $m$ is the total number of noisy speech signals used for classification and $J$ is the total number of the subcategories and $m_j$ and $m_c$ are the numbers of all noisy speech signals present in class $j$ and cluster $c$ respectively\textsuperscript{14}. In this evaluation, the value of $m$ is 720 and $J$ is 4. When the value of $J$ is 1, it denotes the estimated noise level of the noisy speech signal is at 0 dB. Similarly when, the value of $J$ as 2, 3, 4 denotes the estimated noise level is at 5 dB, 10 dB, and 15 dB respectively.

$$Pr(j,c) = \frac{m_{j,c}}{m_c} , Rc(j,c) = \frac{m_{j,c}}{m_j}.$$
\[ F_m(c) = \frac{2 \times Pr(j,c) \times Rc(j,c)}{Pr(j,c) + Rc(j,c)} \quad \ldots (3) \]

**Prediction Speed and Training Time**

To identify how fast the classifiers are working, the Prediction Speed and Training Time measures are used.

**Simulation Results**

The simulations are carried out using the Classification Learner App and Statistics and Machine Learning Toolbox of the MATLAB software in Intel core i5 processor. To increase the accuracy in the classification, the PCA and NCFS methods are used before the classification stage. There are many machine learning-based classification techniques reported in the literature. It has been reported and observed that depending upon the input data, the classification scheme is chosen. In other words, the performance of the classification scheme depends upon the types of input data. For of classification of noisy speech, variety of classification techniques and the performance has been evaluated and it is observed that four classification schemes such as KNN, NB, SVM, DT classifiers are found to be more suitable and yield improved classification accuracy. Based on the above reasons, these four types of classifiers are used and the details are listed in Table 1. Twelve different combinations of feature reduction methods (PCA and NCFS) and four classifiers (KNN, NB, SVM, DT) are developed and simulated for performance evaluation.

Based on the classification accuracy, the weightage of each feature is evaluated using the NCFS method and its bar plot is shown in Figure 2. The higher is the feature weight the more it contributes towards the accuracy of classification.

It is evident from Figure 2 that the features \( f_1, f_9, f_{11}, f_{12}, f_{14}, f_{17} \) are found out to be significant for improving the classification efficiency. All the classifiers are simulated and tested under similar conditions with the fivefold Cross-validation. The CA, Fm, prediction speed and training time of the different classifiers are plotted in Figures 3 and 4. By comparing all the respective evaluation parameters it is observed that the FSVN provides the highest classification accuracy and F-measure in classification, while FDTN gives the highest prediction speed and the WKNP requires the lowest training time.

![Image](image-url)

Table 1 — List of combinations of Feature reduction and Classifiers used in the simulation study

<table>
<thead>
<tr>
<th>SI No</th>
<th>Classifier Type</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Weighted K-Nearest neighbor</td>
<td>WKN</td>
</tr>
<tr>
<td>2</td>
<td>Fine Gaussian Support Vector Machine</td>
<td>FSV</td>
</tr>
<tr>
<td>3</td>
<td>Fine Decision Tree</td>
<td>FDT</td>
</tr>
<tr>
<td>4</td>
<td>Kernel Naive Bayes</td>
<td>KNB</td>
</tr>
<tr>
<td>5</td>
<td>PCA + Weighted K-Nearest Neighbor</td>
<td>WKNP</td>
</tr>
<tr>
<td>6</td>
<td>NCFS + Weighted K-Nearest Neighbor</td>
<td>WKNKN</td>
</tr>
<tr>
<td>7</td>
<td>PCA + Quadratic Support Vector Machine</td>
<td>QSVP</td>
</tr>
<tr>
<td>8</td>
<td>NCFS + Fine Gaussian Support Vector Machine</td>
<td>FSVN</td>
</tr>
<tr>
<td>9</td>
<td>PCA + Fine Decision Tree</td>
<td>FDTN</td>
</tr>
<tr>
<td>10</td>
<td>NCFS + Fine Decision Tree</td>
<td>FDTN</td>
</tr>
<tr>
<td>11</td>
<td>PCA + Kernel Naive Bayes</td>
<td>KNBN</td>
</tr>
<tr>
<td>12</td>
<td>NCFS + Kernel Naive Bayes</td>
<td>KNBN</td>
</tr>
</tbody>
</table>
The confusion matrix obtained from the simulation study shown in Figure 2 for the FSNN algorithm gives the highest classification accuracy.

![Fig. 4 — Comparative analysis of the Classification Accuracy and F-measure of different classifiers](image)

### Conclusion

In the area of speech processing field, the noise level estimation and classification play important roles but it is still not fully explored. This paper has identified important features like Spectral Centroid, Spectral Skewness, Spectral Spread, Spectral Tonal Power Ratio, Entropy and Standard Deviation which are sensitive to the change in the noise level of the speech signal. The performance of four different classifiers is compared with and without the use of PCA and NCFS feature reduction methods. The comparison of classification accuracy, prediction speed and training time reveals that the Fine Gaussian Support Vector Machine with NCFS provides the overall best performance. The classification accuracy found out to be 97% while the F-measure, prediction speed and training time are quite satisfactory. In the future, the outcome of the proposed feature analysis can be combined with the speech enhancement and the speech recognition algorithms to further improve the performance.

### References