FRAME-LEVEL NOISE CLASSIFICATION IN MOBILE ENVIRONMENTS

Khaled El-Maleh[1] *Ara Samouelian*[2] *Peter Kabal*[1]

[1] Dept. Electrical & Computer Engineering
McGill University
Montreal, Quebec H3A 2A7, Canada

[2] School of Elect., Comp. & Telecom. Engineering
University of Wollongong
Wollongong, Australia

Abstract

Background environmental noises degrade the performance of speech-processing systems (e.g. speech coding, speech recognition). By modifying the processing according to the type of background noise, the performance can be enhanced. This requires noise classification. In this paper, four pattern-recognition frameworks have been used to design noise classification algorithms. Classification is done on a frame-by-frame basis (e.g. once every 20 ms). Five commonly encountered noises in mobile telephony (i.e. car, street, babble, factory, and bus) have been considered in our study. Our experimental results show that the Line Spectral Frequencies (LSF's) are robust features in distinguishing the different classes of noises.

1 Introduction

In our daily life, we encounter different types and levels of background acoustical noises (e.g. traffic noise, car noise, office noise etc.). Speech-processing systems (e.g. speech coding, speech recognition, speaker verification) pick up those 'unwanted' signals along with speech. These noise signals result in performance degradation of those systems. For example, the accuracy of a speech recognition device might severely be affected if the level of noise is high and there is a mismatch between training and operation conditions [1]. In speech coding, background noises can be coded with annoying artifacts [2].

Noise classification can be used to reduce the effect of environmental noises on speech processing tasks. As an example, in variable bit rate speech coders, the lowest rate is used to encode background noises in non-active speech periods. As environmental noises vary in texture and dynamics, using one coding scheme has proven to be not adequate for many common types of noises. Noise classification can be used to design natural-quality multi-mode noise coding algorithms. Similarly, multi-mode comfort noise generators can be designed to remedy the noise contrast problem reported in discontinuous transmission-based cellular systems (i.e. GSM). Recently, the issue of background noise is being studied by the ITU-T Study Group 12 (Question 17/12: "Noise aspects in evolving networks" [3]. Noise classification is one of the major parts of this study. In [4], a fuzzy logic noise classifier was designed to distinguish stationary from non-stationary noises at the frame level.

Noise classification has been used in many other applications. For example, in programmable hearing-aid devices, a classification algorithm automatically matches a program mode with the listening environment of the user [5]. In noise monitoring systems, classification of environmental noises is done to help in controlling noise pollution [6].

In this paper, we present the results of our work in designing noise classification algorithms to be used as part of speech-processing systems in mobile environments.

This paper is organized in five parts. Section 2 discusses the feature extraction module and the classification algorithms that have been used for our study. In Section 3, we review the performance evaluation tools we have used. Classification results from different tests of the classification algorithms are given in Section 4. Finally conclusions are presented in Section 5.

2 Frame-Level Noise Classification

2.1 Feature Extraction

The choice of signal features is usually based on *a priori* knowledge of the nature of the signals to be classified. Features that capture the temporal and spectral structure of the input signal are used. Examples of such features are zero crossing rate, root-mean-square energy, critical bands energies, and correlation coefficients. The classifier operates on a frame-by-frame basis using short segments of the signal, e.g. 20 ms.

Linear Prediction (LP) analysis is a major part of many modern speech-processing systems. Transformations of linear prediction coefficients (LPC) (e.g. cepstral, log-area ratio coefficients, line spectral frequencies) have been used successfully in many pattern-recognition problems (e.g. speech recognition, speaker recognition) [7].

We have experimented with different sets of features derived from both the LP coefficients and the LP residual (e.g. residual critical band energies, zero crossing rate). The line spectral frequencies (LSF's) gave the best class separability for the noises we considered. Moreover, a Gaussian fit to each LSF histogram was found to be quite good. Thus, we have selected the LSF's as our features for noise classification. A $10th$ order LP analysis is performed every 20 ms using the autocorrelation method. A Hamming window of length 240 samples is used. The LP coefficients are calculated using the Levinson-Durbin algorithm and then bandwidth expanded using a factor $\gamma = 0.994$. The LP coefficients are then converted to the LSF domain.

2.2 Classification Algorithms

Four pattern-recognition techniques have been chosen for our noise classification problem: Quadratic Gaussian Clas-

sifier (QGC), Least-Square Linear Classifier (LS-LC) [8], Nearest-Neighbor Classifier (NNC) [9], and Decision Tree Classifier (DTC) [10]. A brief description of these classification algorithms is presented below.

A Gaussian classifier is based on the assumption that feature vectors of each class obey a multivariate Gaussian distribution. Estimates of the parameters of the Gaussian PDF of each class (mean and covariance) using the labelled training data are computed. In the classification stage, an input vector is mapped to the class with the largest likelihood. In linear classifiers, a linear discriminant function is optimized to maximize class separability. Least-square optimization algorithm is used to compute the coefficients (weights) of the linear function.

In NN-type classifiers, for each input feature vector, a search is done to find the label of the vector in the dictionary of stored training vectors with the minimum distance. Euclidean distance is commonly used as the metric to measure neighborhood. In k-NN decision rule, the input feature vector is assigned the label most frequently represented among the k nearest patterns in the training dictionary. One major disadvantage of NN classifiers is the need to store large number of training vectors resulting in a large amount of computations. As a remedy to this problem, only prototype vectors from the training data are computed and stored (prototype nearest-neighbor classifier).

A decision tree classifier belongs to the family of machine learning techniques. During the training phase, a set of production rules are generated from the labelled data in the form of a decision tree. The decision tree is then used to classify unlabelled test vectors. The inductive tool used in this paper is an implementation of the C4.5 programs developed by Quinlan [10]. Inductive learning produces decision trees that use the most discriminative features. For more details about decision tree-based classification see [10] and [11].

3 Performance Evaluation

Five commonly encountered noise classes were considered: car, voice babble, street, bus, and factory. A total of 56250 frames (18.75 minutes), equally distributed between the 5 classes, have been used for training. We have recorded street noise (traffic noise, pedestrians walking and talking, and noise from a nearby work area) and bus noise (background music, background speech, bus engine noise, and other external transient noises such as passing cars). The other noises are from the NOISEX-92 database [1]. We have used the Tooldiag pattern recognition software developed by Rauber [12] in designing and testing the QGC, LS-LC, and the k-NN classifiers. The C4.5 inductive learning tool has been used for the DTC.

To measure the discriminating power of the LSF's as features, we estimated the Bayes error rate. A lower bound on the Bayes error rate P_{Bayes} is a function of the asymptotic error rate of the nearest-neighbor decision rule P_{NN} [9], given as

$$P_{Bayes} \geq \frac{M-1}{M}\left(1-\sqrt{1-\frac{M}{M-1}P_{NN}}\right), \qquad (1)$$

where M is the number of classes.

This lower bound will be used as our reference point for the performance evaluation of the different designed classifiers. For each classification algorithm, we used a cross-validation testing methodology to evaluate the classification performance. A 30% of the labeled frames, selected at random are used as test vectors while the remaining vectors were used for classifier training. Five iterations of the Hold-out cross-validation method [8] have been used to compute the empirical error rate for each classifier.

4 Classification Results

In some speech-processing tasks, we need to discriminate speech from noise. For example, a voice activity detection is used in some wireless communications systems to enhance systems capacity and prolong the battery life of portable units. Thus, in this paper we have considered two cases: noise-only classification (5 noise classes), and noise-and-speech classification (5 noise classes and speech class). The classification algorithms were tested using 500 frames (different from the training data) for each class, and with other new noises.

4.1 Noise-only Classification

Table 1 gives the empirical error rate evaluated with the hold-out procedure for the various classifiers. Using Eq. (1) and the empirical error rate of the 1-NN classifier (19.8%), the Bayes error rate was estimated at 10.6%. This means that independent of the classifier structure, the best frame-level classification accuracy for the 5 selected noises (car, street, babble, bus, and factory), and with the 10 LSF's as features is around 89.0%. From Table 1, both the decision tree classifier and the quadratic Gaussian classifier approach that optimal error rate with 11.9% and 13.6% respectively. The linear, and the nearest-neighbor classifiers are less accurate. For the remainder of the paper, we will focus on comparing the performance of the decision tree and the quadratic Gaussian classifiers.

Table 1 Empirical error rate for the different classifiers (noise-only)

Classifier	Error Rate %
Optimal Bayes	10.6
Decision Tree	11.9
Quadratic Gaussian	13.6
3-Nearest Neighbor	17.5
1-Nearest Neighbor	19.8
Linear (least-squares method)	21.9

A detailed presentation of the classification results for each class is given in the form of a classification matrix. Tables 2 and 3 show that the classification accuracy is different for each class. For example, accuracy ranging from 90-100% were obtained for car noise, and factory noise. Street, babble, and bus noises are more often misclassified with error rates ranging from 20-35%. Even though the decision tree classifier has a lower empirical error rate than the Gaussian classifier, the QGC is more robust to new test

vectors. This is due to the parametric nature of the QGC and its ability to model well the LSF's feature vectors.

Table 2 Classification matrix: QGC (noise-only)

	Babble %	Car %	Bus %	Factory %	Street %
Babble	79.8	0.0	12.8	2.0	5.4
Car	0.0	99.6	0.2	0.2	0.0
Bus	8.8	0.0	85.2	2.2	3.8
Factory	1.0	0.0	5.6	93.2	0.2
Street	1.8	0.0	24.8	2.0	71.4

Table 3 Classification matrix: DTC (noise-only)

	Babble %	Car %	Bus %	Factory %	Street %
Babble	71.2	0.0	17.8	3.2	7.8
Car	0.0	100.0	0.0	0.0	0.0
Bus	16.8	0.0	73.6	4.4	5.2
Factory	2.0	0.0	6.4	91.2	0.4
Street	5.2	0.0	25.4	2.8	66.6

4.2 Classification of New Noises

In practical applications of noise classification, the input noise signals are not constrained to belong to one of the 5 pre-selected noise classes. Thus, a 'good' noise classifier should have the ability to map an input feature vector from a new noise class to the closest pre-selected classes. We have tested both the QGC and the DTC on 5 new noise signals (restaurant, shopping mall, sports, subway, and traffic). The results are presented in Tables 4 and 5. It is interesting to observe that both classifiers map the new noises to the noise classes with the same noise events. As an example, a restaurant noise is composed of simultaneous conversations (babble noise), background music, and other ambient noises. Thus, restaurant noise was classified 82.4% as babble noise and 10.8% as bus noise (which has babble, and background music).

Table 4 Classification of new noises: QGC

Noise	Babble %	Car %	Bus %	Factory %	Street %
Restaurant	82.4	0.0	10.8	4.0	2.8
Shop. Mall	52.4	0.0	2.4	0.0	45.2
Sports	22.8	0.0	6.2	0.0	71.0
Subway	53.0	0.0	21.0	4.0	22.0
Traffic	2.7	0.0	15.2	0.0	82.1

4.3 Noise-and-Speech Classification

In Tables 6–8, we present the classification results for the noise-and-speech case. Similar results to the noise-only case were obtained. QGC outperforms DTC in classifying speech and bus noise, with 10% difference in accuracy. Speech signal is 91% accurately discriminated from the noises using the QGC. This suggests that the QGC classifier using LSF's as the features provides robust voice activity detection at the frame level.

Table 5 Classification of new noises: DTC

Noise	Babble %	Car %	Bus %	Factory %	Street %
Restaurant	66.5	0.0	18.7	7.0	7.8
Shop. Mall	26.2	0.0	14.7	0.6	58.5
Sports	19.8	1.5	12.3	9.5	56.9
Subway	48.9	0.0	24.0	2.3	24.8
Traffic	0.8	0.0	4.7	0.0	94.5

Table 6 Empirical error rate for the different classifiers (noise-and-speech)

Classifier	Error Rate %
Optimal Bayes	10.1
Decision Tree	11.6
Quadratic Gaussian	13.6
3-Nearest Neighbor	16.2
1-Nearest Neighbor	18.9
Linear (least-squares method)	33.8

4.4 Classification of Human Speech Like Noise

Human speech-like noise (HSLN) is a kind of babble noise generated by superimposing independent speech signals. HSLN of various number of superpositions (N) (1, 2, 4, ...,1024, 4096) were used in [13] to investigate perceptual discrimination of speech from noise. For example, for low number of superpositions (below 10), the resulting signal will sound like a few speakers' speech. For N between 10 and 200 superpositions, the noise sounds like many speakers in an auditorium (babble-like noise). As N increases, the noise start to sound like stationary Gaussian-like noise (Central Limit Theorem). We have used this set of signals (150 frames each) to test our designed classifiers. The frame-level classification results are shown in Table 9 for the noise-only case and in Table 10 for the noise-and-speech case. Due to the space limitation, we only show the results from the Gaussian classifier. For the noise-only case, the HSLN signals were classified as babble noise most of the time or as bus noise (note that bus noise has babble as one of its noise events). However, for the noise-and-speech case, the results are more interesting. For example, for $N = 1$, the signal is classified as speech (86.8%) and as babble (9.3%). On the other hand, for 128 superpositions, the HSLN signal is classified as babble noise (88.7%) and as speech (5.3%). As N increases, the HSLN signal is classified more as babble than speech. These classification results clearly illustrate the robustness of the designed

Table 7 Classification matrix: QGC (noise-and-speech)

	Speech %	Babble %	Car %	Bus %	Factory %	Street %
Speech	91.0	7.4	0.0	0.8	0.2	0.6
Babble	4.2	76.0	0.0	12.4	2.0	5.4
Car	0.2	0.0	99.6	0.2	0.0	0.0
Bus	2.4	8.0	0.0	84.0	2.2	3.4
Factory	0.2	1.0	0.0	5.6	93.0	0.2
Street	0.2	1.8	0.0	24.6	2.0	71.4

Table 8 Classification matrix: DTC (noise-and-speech)

	Speech %	Babble %	Car %	Bus %	Factory %	Street %
Speech	81.6	14.4	0.0	1.8	0.8	1.4
Babble	5.0	70.6	0.0	14.4	2.8	7.2
Car	1.4	0.0	98.4	0.0	0.2	0.0
Bus	2.6	14.4	0.0	74.0	3.4	5.6
Factory	0.6	1.2	0.0	6.4	91.4	0.4
Street	0.6	4.4	0.0	26.2	2.6	66.2

Gaussian classifier. Similar results were obtained for the decision tree classifier.

Table 9 Classification of HSLN signals: QGC (noise-only)

N	Car %	Factory %	Street %	Bus %	Babble %
1	0.7	6.6	0.0	0.7	92.0
4	0.0	0.0	0.7	4.6	94.7
8	0.0	0.0	3.3	11.3	85.4
16	0.0	0.0	3.3	7.3	89.4
32	0.0	0.0	2.0	4.0	94.0
128	0.0	0.0	0.7	5.3	94.0
512	0.0	0.0	3.3	8.0	88.7
1024	0.0	0.0	5.3	3.3	91.4
4096	0.0	0.0	4.6	10.6	84.8

Table 10 Classification of HSLN signals: QGC (noise-and-speech

N	Speech %	Car %	Factory %	Street %	Bus %	Babble %
1	86.8	0.7	3.4	0.0	0.0	9.3
4	81.5	0.0	0.0	0.7	1.3	16.6
8	72.9	0.0	0.0	3.3	5.9	17.9
16	51.0	0.0	0.0	3.3	4.0	41.7
32	29.8	0.0	0.0	2.0	3.3	64.9
128	5.3	0.0	0.0	0.7	5.3	88.7
512	0.7	0.0	0.0	3.3	8.0	88.0
1024	1.3	0.0	0.0	5.3	3.3	90.1
4096	0.7	0.0	0.0	4.6	10.6	84.1

5 Conclusion

Frame-level noise classification results have been presented using four pattern-recognition frameworks. The line spectral frequencies have been used as the features. The quadratic Gaussian classifier outperforms the other classifiers tested. The accuracy can be improved substantially by postprocessing the temporal sequence of decisions (for instance with a Viterbi type algorithm), however this comes at the expense of further delay.

References

[1] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," 1992. `http://spib.rice.edu/spib/select`.

[2] T. Wigren, A. Bergstrom, S. Harrysson, F. Jansson, and H. Nilsson, "Improvements of background sound coding in linear predictive speech coders," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Detroit, MI), pp. 25–29, May 1995.

[3] ITU-T, Geneva, *COM 12-1-E- List and wording of questions allocated to Study Group 12 for study during the 1997–2000 study period*, Feb. 1997.

[4] ITU-T, Geneva, *Delayed 11-E (WP 3/12)-Background noise classification in mobile environments*, Mar. 1997.

[5] J. M. Kates, "Classification of background noises for hearing-aid applications," *J. Acoust. Soc. Am.*, vol. 97, pp. 461–470, Jan. 1995.

[6] C. Couvreur and Y. Bresle, "A statistical pattern recognition framework for noise recognition in an intelligent noise monitoring system," *in Proc. EURONOISE'95* (Lyon, France), pp. 1007–1012, Mar. 1995.

[7] C.-S. Liu and M.-T. Lin, "Study of line spectrum pair frequencies for speaker recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Albuquerque, NM), pp. 277–280, Apr. 1990.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[9] T. M. Cover and P. E. Hart, "Nearest-neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. 13, pp. 21–27, Jan. 1967.

[10] J. R. Quinlan, *C4.5. Programs for Machine Learning, Morgan Kaufmann Series in Machine Learning*. Morgan Kaufmann Publisher, 1993.

[11] A. Samouelian, "Frame-level phoneme classification using inductive inference," *Computer Speech and Language*, no. 11, pp. 161–186, 1997.

[12] T. Rauber, *Inductive Pattern Classification: Methods-Features-Sensors*. PhD thesis, Universidade Nova de Lisboa, 1994. `http://www.inf.ufes.br/~thomas`.

[13] D. Kobayashi, S. Kajita, K. Takeda, and F. Itakura, "Extracting speech features from human speech like noise," *Proc. Int. Conf. on Spoken Language Processing*, pp. 418–421, Oct. 1996.