A method for recognition of coexisting environmental sound sources based on the Fisher's linear discriminant classifier

Ester Creixell[1], Karim Haddad[2], Wookeun Song[3], Shashank Chauhan[4] and Xavier Valero.[5]

[1] Danmarks Tekniske Universitet

Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark

[2 3 4] Brüel & Kjær Sound and Vibration Measurement A/S

Skodsborgvej 307, 2850 Nærum, Denmark

[5] La Salle - Universitat Ramon Llull

Quatre Camins 30, 08022 Barcelona, Spain

ABSTRACT

A method for sound recognition of coexisting environmental noise sources by applying pattern recognition techniques is developed. The investigated technique could benefit several areas of application, such as noise impact assessment, acoustic pollution mitigation and soundscape characterization. This study distinguishes from other investigations by focusing on cases where the noise sources appear mixed (i.e., several noise sources might be present at the same time in one location), which is a more realistic and frequent situation in cities than a single sound source without other interfering noises. The identification and, furthermore, the estimation of the contribution of each source to the overall level is one important goal in the current investigation, which would improve environmental noise assessment in complex situations. The method for recognizing the noise sources in adverse conditions is based on the Fisher's Linear Discriminant classifier, and estimates noise source contributions based on a distance measure of vector projections. The method is able to identify mixed sources in 96% of the 27 tested signals and to correlate the contribution of the individual sources with their sound pressure level. The results obtained from tests in real city environments show an accurate performance in the description of the sound scenarios.

1. INTRODUCTION

Environmental noise recognition has several areas of application, yet an important task in which it can contribute is that of mapping environmental sounds in the city environments, which is required by the Environmental Noise Directive (END) [1]. Environmental noise may refer to a wide variety of sounds, from industry to traffic noise or nature sounds. Unfortunately, sound environment in cities is dominated by unwanted noises, which may decrease the quality of life of the population or even become harmful for health. This claims the need for a powerful tool that contributes to ease the task of

[1] s111473@student.dtu.dk
[2] Karim.Haddad@bksv.com
[3] Woo-Keun.Song@bksv.com
[4] Shashank.Chauhan@bksv.com
[5] xvalero@salle.url.edu

noise mapping and sound source characterization. Moreover, environmental noise recognition can be applied in fields like noise control, civil engineering, road planning, acoustic pollution mitigation, security surveillance systems or soundscape characterization (which could be used, for example, in hearing-aids devices for deaf people).

The application of sound recognition techniques to environmental noise has been studied for more than 20 years, leading to great technological advances and high recognition rates in controlled recordings. A typical pattern recognition approach is followed for sound recognition in this paper. This approach consists of two main steps: in the first step a noise sample is analyzed to extract characteristic features, and in the second step the sample is classified according to patterns found in the features. The second step can usually be performed after the classifier has been through a training phase.

Several previous works related to environmental sound recognition can be found in the literature. In Cowling and Sitte [2] an exhaustive review of the most important features and classifiers for non-speech recognition is done, and in Mitrovic [3] the best techniques for speech recognition are studied and applied for different kinds of environmental noises to evaluate the results. As a conclusion from the features tested, the author points at Linear Predictive Coding (LPC) and two kinds of Cepstral Coefficients, Bark Frequency Cepstral Coefficients (BFCC) and Mel Frequency Cepstral Coefficients (MFCC) as the highest discriminative for environmental sounds. In Rodeia [6] MFCC is also chosen for environmental sound discrimination.

In the study by Hansen [4], the features MFCC, Linear Predictive Cepstral Coefficients (LPCC) and Perceptual Linear Predictors (PLP) are tested for environmental sound recognition. PLP yields high recognition rates (comparable to those obtained with MFCC), while LPCC does not achieve such good results. These three features are also tested in Valero and Alias [5] among others, and MFCC is shown to outperform the other two.

As far as classifiers are concerned, k Nearest Neighbors (k-NN) is a simple method that gives good results according to several studies, such as in Mitrovic [3], Valero and Alias [5] and Rodeia [6]. In the study by Sobreira et al. [7] the classifier FLD (Fisher's Linear Discriminant) is used for classifying traffic noise of cars, trucks and motorcycles, and proven to give better results than kNN. In Valero and Alias [5], Rodeia [6] and Ntalampiras et al. [8], other classifiers such as SVM (Support Vector Machines), GMM (Gaussian Mixture Models) and HMM (Hidden Markov Model) are also shown to outperform kNN for classifying environmental noise of different kinds.

In this investigation, the features MFCC, PLP and LPCC are chosen to be tested as the previous works show they obtain the best results. For classification, kNN, GMM and FLD are compared as they also show good performance.

The methods studied in the past years were intended to distinguish between different sound sources [2], however, the current investigations deal with situations closer to reality, i.e. cases where the signal to noise ratio is low, identification of sound sources independently of the attenuation with distance, or situations where the target sound sources appear mixed. This investigation focuses on the latter problem, the main goal being identification and, furthermore, estimation of contribution of each source to the overall level.

The paper is organized as follows. Section 2 introduces the theoretical approach of the proposed solution based on the FLD. Section 3 describes the different experimental setups used for testing the proposed recognition system, Section 4 presents results for single source recognition and Section 5 shows the results of the tests with artificially-mixed sources and real city noise recordings containing a mixture of sound sources.

2. PROPOSED SOLUTION

In this section, a method to detect and quantify noise originated by two or more different sound sources out of a recording is developed based on the FLD classifier. The FLD used in a classical recognition system would classify an input sample into one of the predefined classes, therefore the response would be unique even if the sample actually contained noise from different sources. The objective of this method is to be able to detect the presence of two or more simultaneous noise sources and identify them.

2.1 Fisher Linear Discriminant

The principle of the FLD is to map a set of n-dimensional feature vectors that correspond to two different classes into a hyperplane in such a way that the projections belonging to different classes are

maximally separable. Mathematical procedures to achieve this can be found in Ye et al. [9]. The projections can then be separated by another hyperplane, called the FLD. If more than two classes must be classified, a discriminant for each class is calculated. In Figure 1, an example with two-dimensional feature vectors is shown. There are three classes to be separated, therefore three FLDs are calculated, which is done by considering the class of interest against all the other classes in each case.

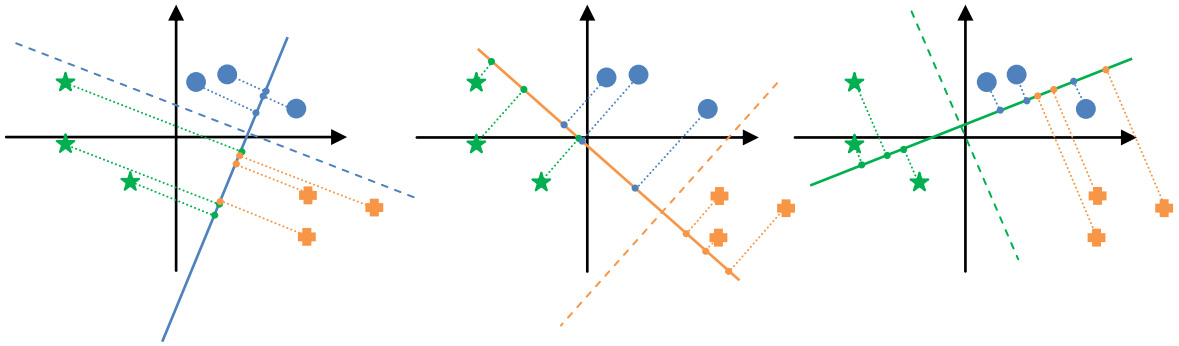

Figure 1 – FLDs (dashed lines) for 3 class separation in a 2-dimensional case.

The FLDs are calculated in the preliminary training phase. When a new sample is to be classified, the distances to the FLDs are calculated and the sample is assigned to the class with longer positive distance. An example is shown in Figure 2, where the red cross represents a new sample that would be classified as "Car" given that $d2$ is the longest distance.

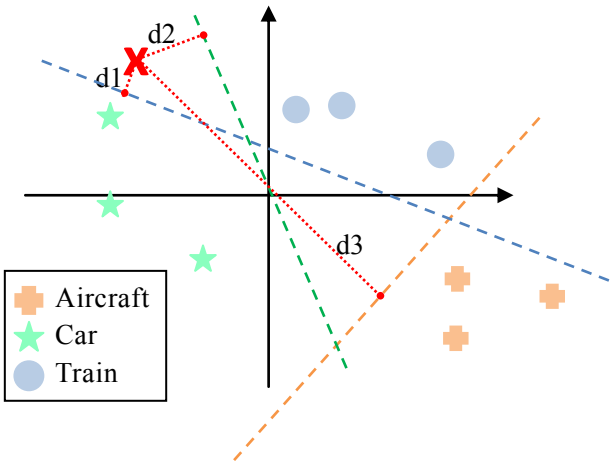

Figure 2 - Classification for a test sample

2.2 Mixed sources identification

The fact that $d1$ is also positive means that the sample is also in the "Train" class space. The hypothesis of the new method in such a case is that the analyzed sound sample contains sound from both car and train sources. In this case, a percentage of belonging to each of the classes can be calculated as

$$\textit{belonging to class } i\ (\%) = \frac{d_i}{\sum_{n=1}^{N}(d_n > 0)} \tag{1}$$

Where i denotes one of the classes, d_i denotes the distance of the sample to the discriminant of class i, and the summation in the denominator includes all the positive distances (in the example of Figure 2, $d3$ would be excluded).

As a result, the new system output is a percentage of belonging of each audio frame to each of the classes, instead of one single label. A comparison of the output for a given input signal with mixed train and car noise is shown in Figure 3.

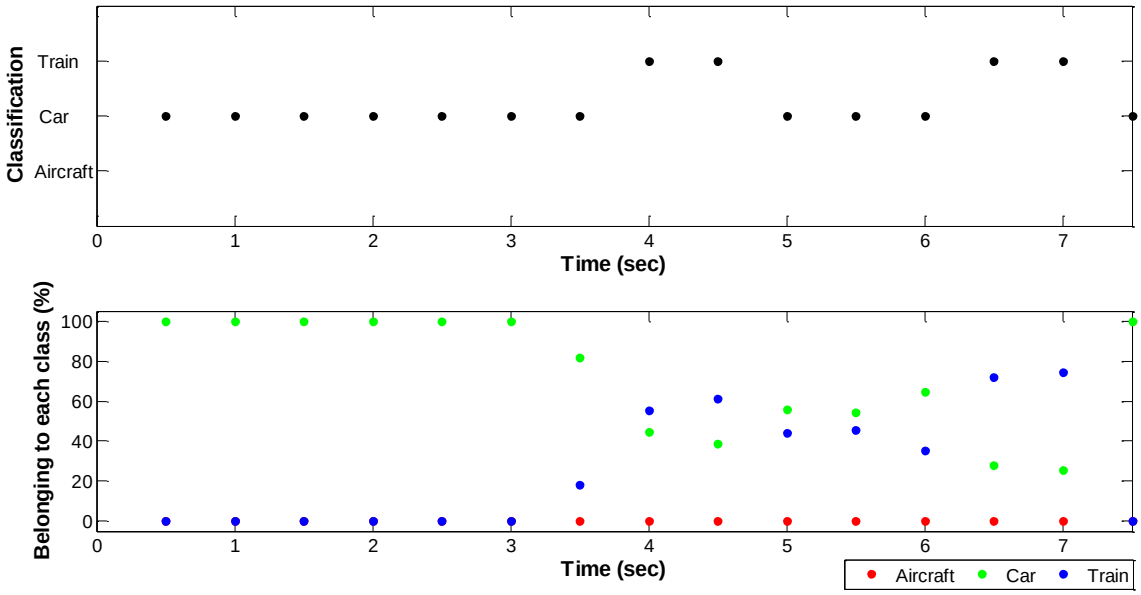

Figure 3 - Top: Classification result of the mixed input sample.

Bottom: Percentage of belonging to each class calculated according to eq.(1).

The system is configured to produce an output every 0.5 s as detailed in section 4. The top plot shows that 4 time segments are classified as "Train" and 11 are classified as "Car", as in a typical FLD result. However, in the bottom plot, the percentages show that the 6 first time segments are 100% "Car" while the rest are in the positive side for the two classes, and in which percentage they are bound to be "Train" or "Car".

3. EXPERIMENTAL SETUP

3.1 Database

A database composed of sound samples from car, train and aircraft noise is used for the tests. The recordings can be divided in two sets: set 1 contains single source recordings and set 2 contains mixed source recordings. Set 1 is used for the preliminary tests described in section 4, where it is divided in two subsets: training and test. In the experiments of section 5, the whole set 1 is used for training the recognition system. Table 1 shows the composition of this set, as a summation of the times of all samples, which were recorded in different locations.

Table 1 – Composition of Set 1 – single source recordings (seconds)

	Aircraft	Car	Train	Total
Training	235	165	79	479
Test	234	167	79	480

Training set 2 contains city noise recordings which have been made in places where different sound sources can be heard at the same time. Specifically, two kinds of acoustic environments were chosen: locations where cars and trains can be heard and locations where cars and aircrafts are present. Those will be used as test samples in the experiments in section 5.2.

All recordings were made using the sound level meter type 2250 from Bruel & Kjær. The sampling rate for recording is 24 kHz.

3.2 Test setups

The hypothesis is tested by means of artificial mixtures. For this purpose, 3 samples of each class are selected from the database set 1, which contains single noise sources. Each sample is mixed with one sample of the other classes, resulting in 9 different mixtures. The criteria used for the selection is that every independent source must obtain more than a 90% of belonging to its class when analyzed individually. In this way, the results can be interpreted based on the mixed source method, as it is assured that the individual source classification is working satisfactorily. The individual sound sources

are scaled so as both have the same contribution to the artificial mixture, in terms of RMS. These mixtures are used as test samples in section 5.1.

To ease the visualization of the results, a total percentage of belonging is plotted for each input signal. This is calculated by adding the percentages for each class from all the time segments and dividing them by the number of segments. For the example of Figure 3, the total percentages would be 73% car, 27% train and 0% aircraft.

Once the hypothesis is tested in a controlled setting, and in order to see the effectiveness of the method when applied in a real situation, the recordings containing mixed sources from real city environments are tested.

4. RESULTS FOR SINGLE SOURCE RECOGNITION

As a preliminary stage to the identification of mixed sources, the selected methods - the features MFCC, PLP and LPCC and the classifiers kNN, GMM and FLD - were tested with the recordings in set 1 from section 3. The structure of the recognition system is shown in Figure 4.

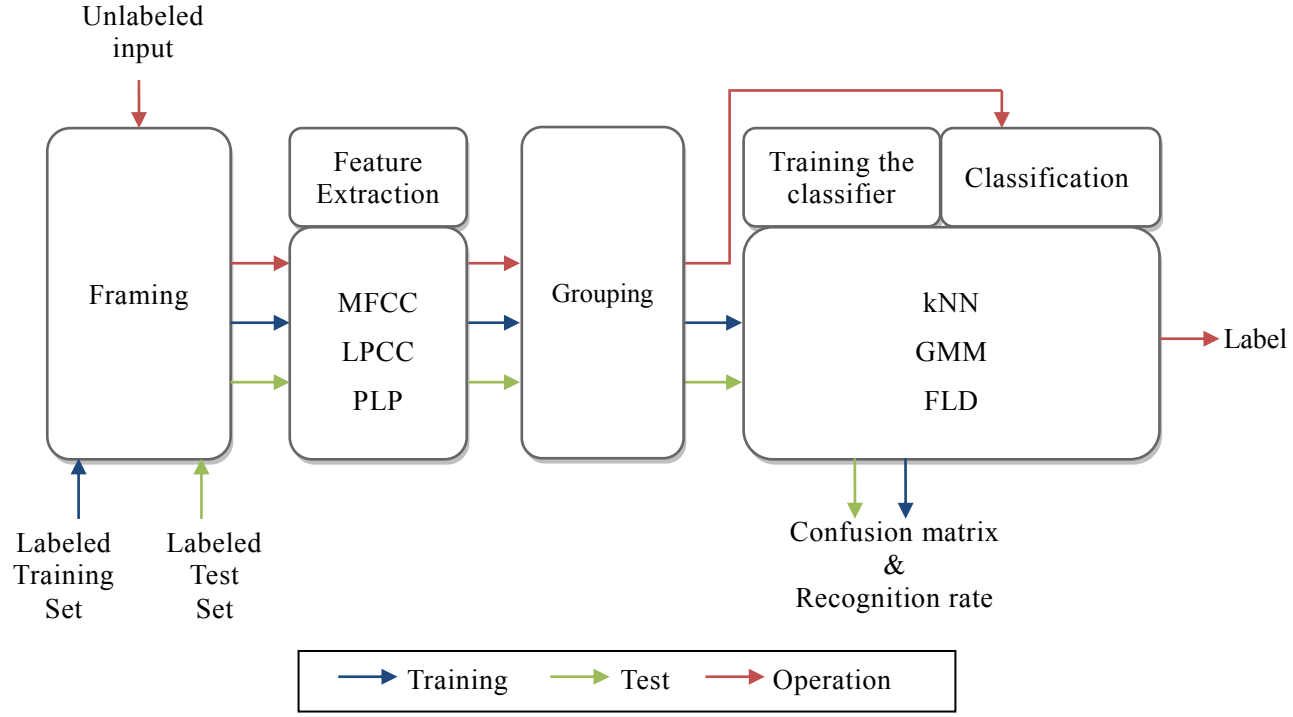

Figure 4 - Diagram of the recognition system

The feature extraction and classifier blocks are the main components of a recognition system. Additionally, it includes a framing block (i.e. the input signal is windowed in smaller segments, namely frames) and a grouping block, where feature vectors are averaged over several frames to take into account the time evolution of the signal.

The system works in two phases illustrated by the blue and red arrows in Figure 4. In a preliminary phase, the training process of the classifier takes place: the system is fed with labeled sound samples, feature vectors are extracted and the classifier relates them to their corresponding classes. After that, the trained classifier is ready to identify unknown samples in the operation phase, where each input frame is assigned to a class.

Yet another phase illustrated by the green arrow can be used to test the system. In this case the trained system is fed with known samples, which are classified by the system as it would do with unlabeled input and finally the system response is compared with the real answer. In this way, a percentage of correct identifications can be calculated (i.e. recognition rate).

The recognition rates obtained from different feature-classifier combinations can be seen in Table 2. Further details on the parameters used for the tests can be found in Creixell [10]. The results showed that FLD is the classifier with the best performance when MFCC and PLP are used, with a recognition rate of about 90% in both cases. Based on these results, the FLD is chosen to develop the method for identifying mixed sources.

Table 2 – Recognition rate for different feature-classifier combinations

Classifiers \ Features	MFCC	LPCC	PLP
FLD	90,7 %	72,1 %	90,6 %
GMM	88,7 %	56,1 %	85,0 %
kNN	82,3 %	76,9 %	87,0 %

5. RESULTS FOR MIXTURE OF SOURCES

5.1 Validation with artificially-mixed noise signals

The 9 mixed signals described in section 3.2 have been analyzed by the system using two feature extraction methods: MFCC and PLP. The results are shown in Figure 5 and Figure 6. The samples named "aircar" denote the mixture of an aircraft and a car signal, the samples named "airtrain" denote the mixture of an aircraft and a train signal, and the samples named "cartrain" denote the mixture of a car and a train signal.

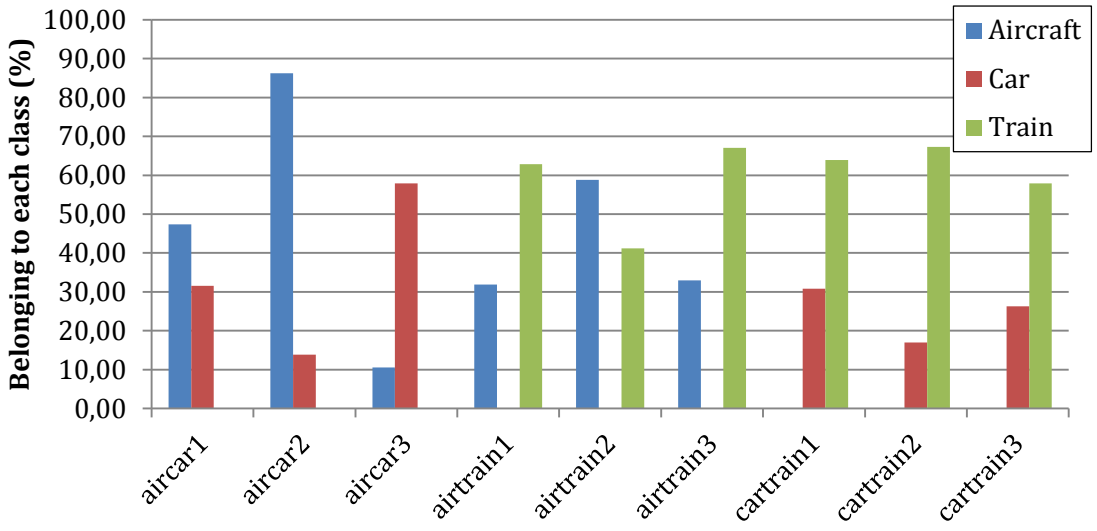

Figure 5 - Percentage of belonging of the mixtures.

Parameters: 8 MFCC coefficients, groups of 5 feature vectors, 100 ms frames, FLD.

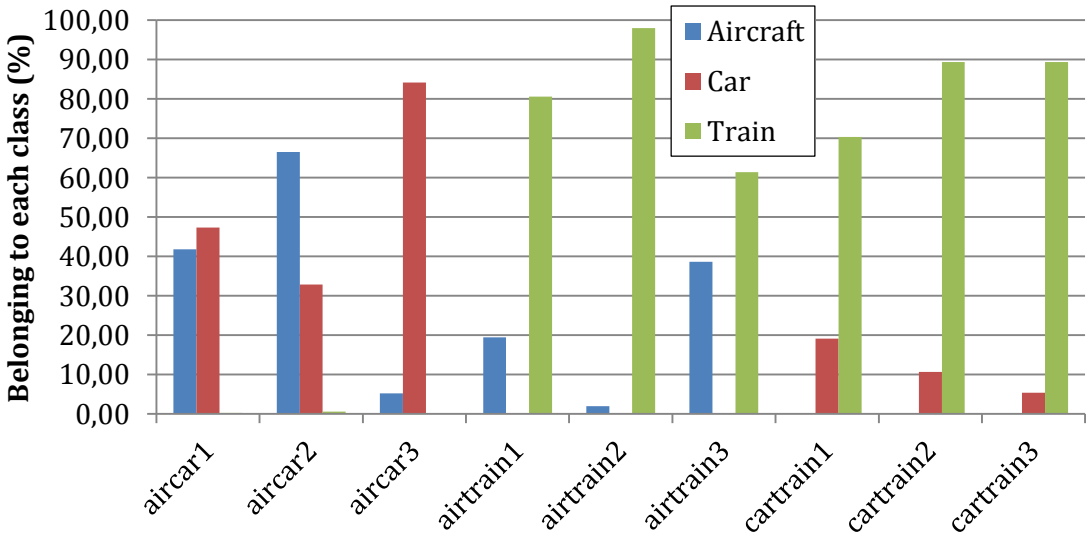

Figure 6 - Percentage of belonging of the mixtures.

Parameters: 8 PLP coefficients, groups of 5 feature vectors, 100ms frames, FLD.

In all cases, the two expected classes are detected, since they present percentages above 0%. Moreover, the unexpected class is never detected by the system (percentages below 1%), meaning that the individual sources are detected successfully. It should be pointed that the PLP shows a tendency to emphasize the "Train" class over the rest, since it gives higher percentages to it in all cases. Therefore, MFCC is selected for the forthcoming experiments.

It could be expected that a 50% chance of belonging to each of the two classes should be obtained given that the two signals that compose each mixture have the same RMS value, however, this is not true for each frame but for the whole signal, thus the time evolution of the signals has an important role.

Still, a relation between the energy of each signal and the assigned percentage can be established. Another series of mixtures is created by picking one train sample and one aircraft sample. They are scaled so as to have the same RMS, and then mixed with different proportions, meaning that the aircraft sample is weighted by a coefficient that ranges from 0 to 2 in steps of 0.2, while the train signal remains constant. Therefore, when the coefficient is 1 both signals have the same RMS. The signals are then processed by the recognition system and the percentages of belonging to each class are obtained for each mixture. A relation between the RMS of the aircraft signal over the total and the percentage obtained for the class "Aircraft" is shown in Figure 7.

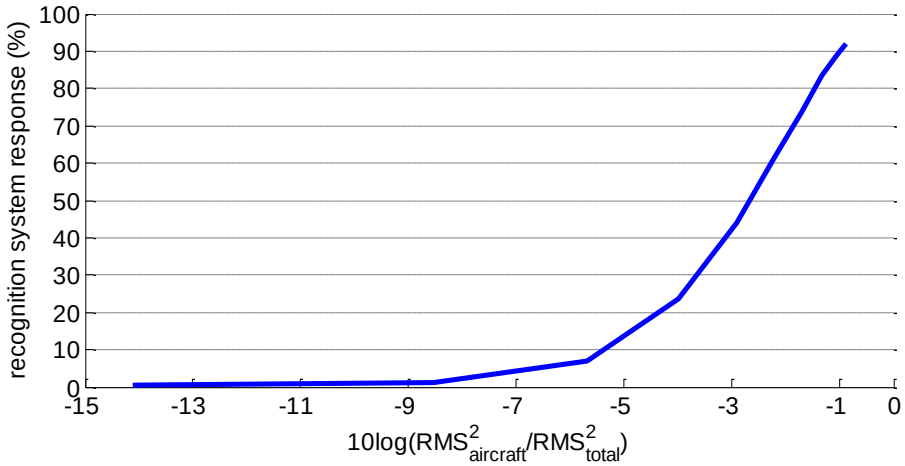

Figure 7 - Percentage of aircraft detected in relation to the proportion of aircraft in the mixture. Parameters: 8 MFCC coefficients, groups of 5 feature vectors, 100ms frames, FLD.

The curve shows that the percentage of belonging to each class given by the recognition system changes according to the proportions of the mixture. The more energy the aircraft signal has in the mixture, the higher the percentage given to its class is. Identical procedures done with mixtures from other classes led to curves with similar shapes, as well as the same experiment done using PLP instead of MFCC.

This proves that a relation can be established between the percentage calculated and the ratio between the source energy and the total energy. Therefore using FLD in combination with MFCC or PLP is a satisfactory method to describe soundscapes with mixed sources.

5.2 Experiments with real environmental noise mixtures

The system is tested for real mixed source recordings in this section. As mentioned above, the selected feature extraction method is MFCC.

A situation where cars and trains can be heard is easy to find in a city, as there are several places where railways and highways meet. One of these places can be seen in Figure 8. Measurements were made in two different locations, indicated by the signs "Loc 1" and "Loc 2". It is easy to notice that in "Loc 1" the railway is closer than the highway, therefore, when a train passes by, its sound level will be higher than that from the cars. On the other hand, in "Loc 2" the highway is closer than the railway, and also a secondary road is very close, therefore the car noise is expected to be louder.

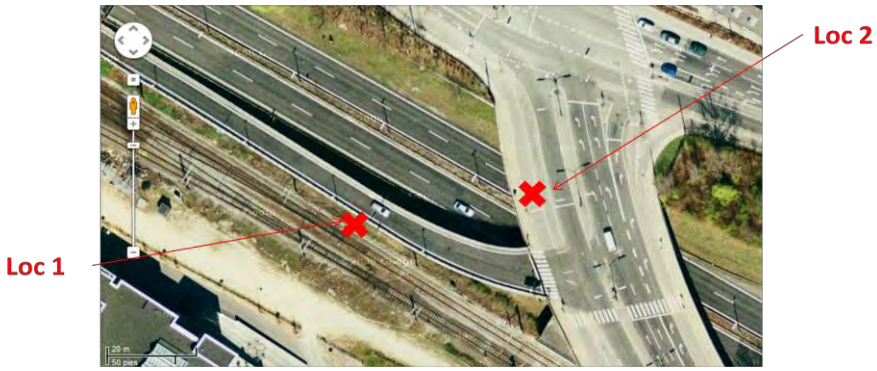

Figure 8 – A map of the location of the measurements.

A result from a recording in "Loc 1" is shown in Figure 9. The recording is composed of background car noise from the highway and a train passing by from second 5 to 11, as indicated top plot in the figure in red. The "Classification" plot in the middle part of the figure shows the system response for its classical behavior in which only 1 class per each group of frames can be the answer. The bottom plot shows the results of the method to detect mixed sources by means of the percentage of belonging of each group of frames to each class. In the first 4 s and from 12 s to the end, the percentages for the class "Car" are very high, while between 5 s and 11 s the percentages for the class "Train" are almost 100%. In the transition periods, the percentages are close to 50%. Therefore the evolution is very well described.

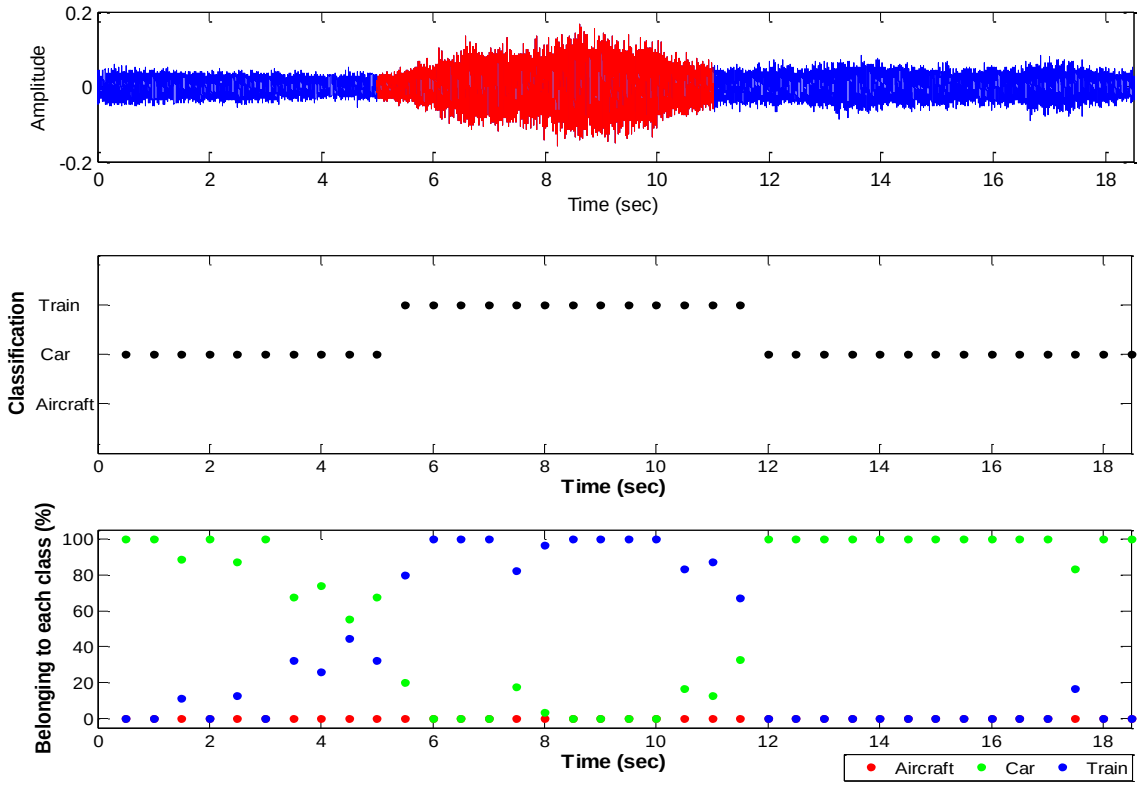

Figure 9 – Recording in "Loc 1". Top: Audio signal waveform. Middle: Response of the single-source recognition system. Bottom: Response of the mixed-source recognition system.

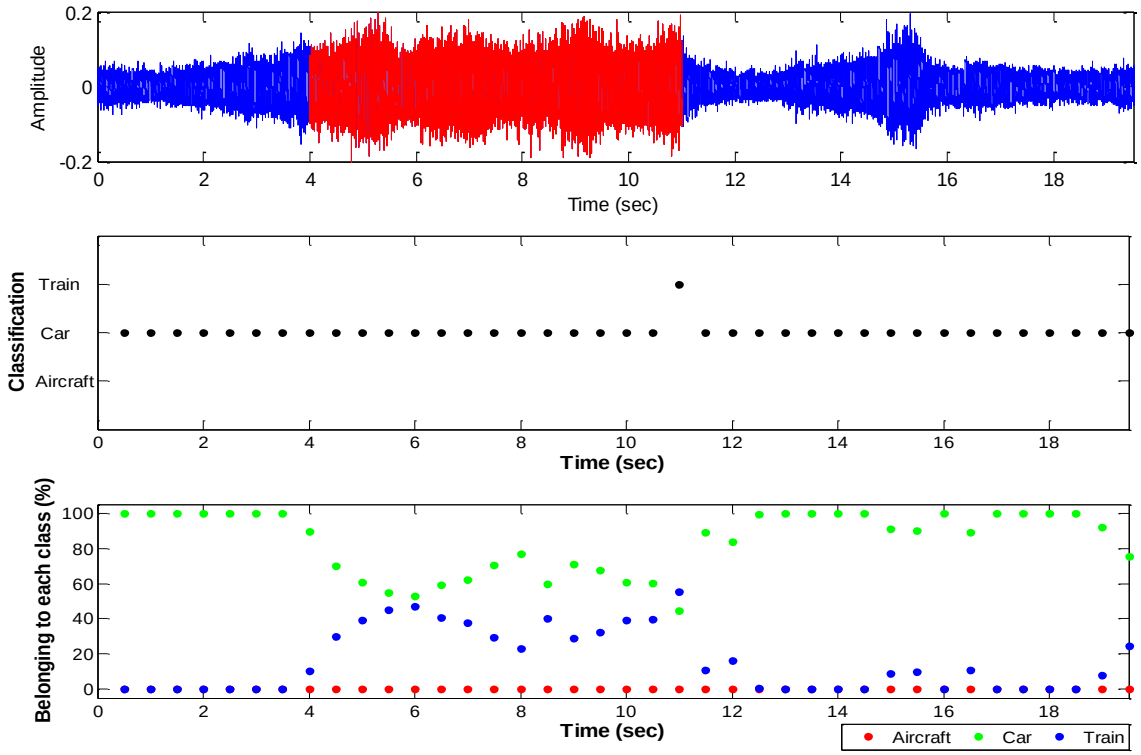

Figure 10 - Recording in "Loc 2". Top: Audio signal waveform. Middle: Response of the single-source recognition system. Bottom: Response of the mixed-source recognition system.

The results from a recording made in "Loc 2" are shown in Figure 10. When listening to the recording, cars can be heard during the whole time of the recording in the foreground, while the train is heard between seconds 6 and 11 in the background, as indicated by the top plot in Figure 10. The percentages for the class train are higher between 6 s and 11 s than for the rest, which corresponds with the subjective perception. In the first 4 s and after from 12 s to the end no train can be heard, therefore the percentage of 0% assigned to the class "Train" in these periods is an accurate description as well.

This example shows an important utility of this method. In Figure 10 the middle plot shows that all the responses except for one would be "Car" with the classical single-source recognition system. Therefore, in this case, if no mixed source detection was used, the results would show no sign of a train passing by; however, the new method detected the presence of both train and car noise and showed how each source contributes to the mixture.

Further tests were performed using recordings from other locations with similar characteristics, and from locations where aircraft and car noises were present simultaneously, which led to similar results and correlation between the response of the system and the subjective perception.

6. CONCLUSIONS

This paper has addressed the problem of environmental sound recognition in situations where the sound sources appear mixed. The proposed technique provided a possibility of detecting the mixture of sources and the contribution of each source to the overall sound pressure level.

A method based on FLD has been introduced to quantify the percentage of belonging to each class by the ratio between the sum of all positive distances and the positive distance of the class of interest. The method has been tested using artificially mixed sources, which are combinations of single source recordings, and has yielded successful detection of individual sources in mixtures, especially with MFCC, yet with PLP the results have also been satisfactory.

Finally, the system has been tested with real recordings. For this phase, only MFCC has been used, given its better performance in the previous experiments. The results obtained are encouraging; the time evolution of the output percentages of belonging to each class are well correlated with the subjective perception that one has from the recordings. The fact that the samples are well recognized even though only single source recordings taken in different locations and times are used for training the system shows its high robustness.

Despite the fact that the system was able to detect the presence of noise sources in a set of mixtures tested in the study, the proposed method needs to be tested on a larger database of samples to generalize the findings.

REFERENCES

[1] X. Valero, F. Alías, S. Kephalopoulos and M. Paviotti, "Pattern recognition and separation of road noise sources by means of ACF, MFCC and probability density estimation," in Proc. Euronoise'09 (Edimburgh, UK, 2009).

[2] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," Pattern Recognition Letters, vol. 24, no. 15, p. 2895–2907 (2003).

[3] D. Mitrovic, "Discrimination and Retrieval of Environmental Sounds," Master Thesis, Technische Universität Wien (2005).

[4] T. H. Hansen, "Classification of Environmental Sounds. Pattern Recognition. Report 2 for bachelor internship.," Technical University of Denmark (2012).

[5] X. Valero, F. Alias, "Hierarchical Classification of Environmental Noise Sources Considering the Acoustic Signature of Vehicle Pass-Bys", Archives of Acoustics, vol. 37, no. 4, pp. 423-434 (2012).

[6] J. Rodeia, "Analysis and recognition of similar environmental sounds," M.Sc. Thesis, Universidade Nova de Lisboa (2009).

[7] M.Sobreira Seoane, A.Rodriguez Molares, J.L.Alba Castro, "Automatic classification of traffic noise", in Proc. Acoustics '08 (Paris, France, 2008).

[8] S. Ntalampiras, I. Potamitis, N. Fakotakis, "Automatic Recognition of Urban Environmental Sound Events", Proc. International Association for Pattern Recognition Workshop on Cognitive Information Processing (2008).

[9] Q. Ye, C. X. Zhao, H. F. Zhang, X. B. Chen, "Recursive ''concave–convex' 'Fisher Linear Discriminant with applications to face, handwritten digit and terrain recognition," Pattern Recognition, vol. 45, no. 1, p. 54–65 (2012).

[10] E. Creixell, "Sound Recognition Techniques: Application to city noise", B.Sc. Thesis, La Salle - Universitat Ramon Llull (2012).