inter.noise 2004
Prague • Czech Republic • August 22–25

The 33rd International Congress and Exposition
on Noise Control Engineering

Automatic Noise Recognition in Urban Environments Based on Artificial Neural Networks and Hidden Markov Models

L. Couvreur[a], M. Laniray[b]

[a]MULTITEL asbl, Avenue Copernic 1, B-7000 Mons, Belgium
[b]01dB Acoustics & Vibrations, Chemin des Ormeaux 200, F-69578, Limonest, France

[a]couvreur@multitel.be;[b]marc.laniray@01db.com

Abstract [133] In this paper, we present an Automatic Noise Recognition (ANR) system that has been developed for detecting and identifying noise acoustic events in urban environments. The proposed system is based on Artificial Neural Networks coupled with Hidden Markov Models. We consider two types of noise events that are generally seen as a nuisance in urban areas, namely scooter engines and horn signals. A sizeable database has been built up and counts up to 1,000 sound samples, from which the ANR system has been developed and tested. Results show good performances as for both detection and identification of noise events.

1 INTRODUCTION

For the past few years, there has been an increasing demand for automatic systems that are able to monitor the acoustic activity in sensitive areas, *e.g.*, in the vicinity of airports or in the center of cities. For example, the production of noise maps has received an increasing interest for the last few years, notably in Europe according to directives on environmental noise [1]. In sensitive areas, the acoustic activity is relatively constant, the so-called background noise level, then it can vary drastically when a noise event happens. We define a noise event as an unexpected increase of the acoustic level, which is generally perceived as annoying. Though it is important to be able to detect noise events and measure their intensity and duration in order to characterize their frequency and their annoyance, one can desire to identify also the source of the noise event. This problem is known as Automatic Noise Recognition (ANR).

In this communication, we present an ANR system that is based on Artificial Neural Networks (ANN) coupled with Hidden Markov Models (HMM). In section 2, we describe the proposed system. In section 3, results are reported for recognition experiments of urban noises. The noise audio database is described, the experimental setup is detailed and results are compared for various configurations. Conclusions and perspectives for future work are given in section 4.

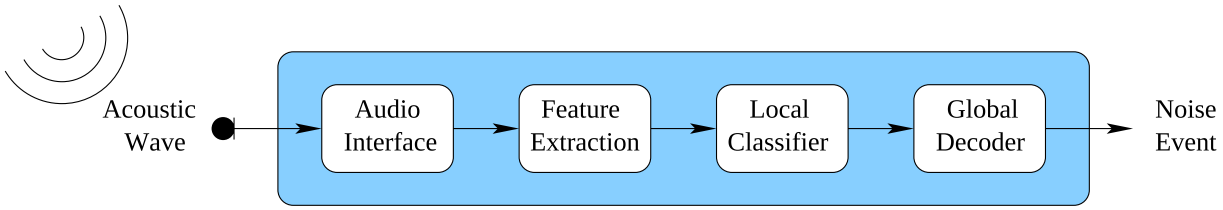

Figure 1: *Block diagram of an Automatic Noise Recognition (ANR) system.*

2 PROPOSED METHOD

The proposed ANR system is depicted in figure 1. It consists of several modules that analyze an acoustic wave, and detect and identify possible noise events. These modules are namely the audio interface, the feature extraction, the local classifier and the global decoder.

2.1 Audio Interface

The acoustic wave that impinges on the noise monitoring system is first converted into an electrical signal by a microphone. Next, an analog-to-digital conversion is applied in order to transform the microphone signal into a form workable by a computer. That is, a digital signal is obtained by discretizing the microphone signal both in time (sampling) and amplitude (quantification). In this work, the sampling frequency F_s is set equal to 44.1 kHz and samples are quantized linearly as 16-bit integers.

2.2 Feature Extraction

Basic Algorithm: Perceptual Linear Predictive

Recognition of a noise event directly from the recorded time signal is a hard problem. First, this representation of the acoustic information is weakly discriminant. Figure 2 shows the time signals (upper plots) for two noise events, namely a scooter pass-by and a horn signal. They look quite similar and are hard to discriminate. Besides, time signals are highly variable, that is, they changes drastically from a noise event to another of the same nature. Consequently, it is difficult to estimate accurately a statistical model (to be used afterwards during recognition) of this representation for every type of noise event. The feature extraction module aims at deriving a set of features from the time signal that better characterize noise events and ease their detection and identification. The proposed method consists in performing a time-frequency analysis, *i.e.* to represent the time evolution of the power spectral envelope in a compact form, by applying the following steps to the time signal.

The very first step consists in blocking the time signal into successive, possibly overlapping, frames. The choice of the frame length W_l results from a tradeoff between signal stationary and frequency resolution, while the frame shift W_s depends on time resolution. In this work, we set $W_l = 60$ ms and $W_s = 20$ ms.

The second step consists in computing the power spectrum for every frame. This is typically done by weighting the frame signal with a Hamming window, zero-padding the windowed signal to the next power of two, computing its Fast Fourier Transform (FFT) [2] and taking its squared magnitude. In figure 2, we show the power spectrograms (lower plots) for a scooter event and a horn event. Power spectrograms are obtained by stacking together the vectors formed with the power spectral coefficients. We clearly observe typical patterns that differentiate the two noise events and allow to recognize them. However, these sets of features are still badly conditioned for recognition purpose. It is required to reduce their dimensionality and

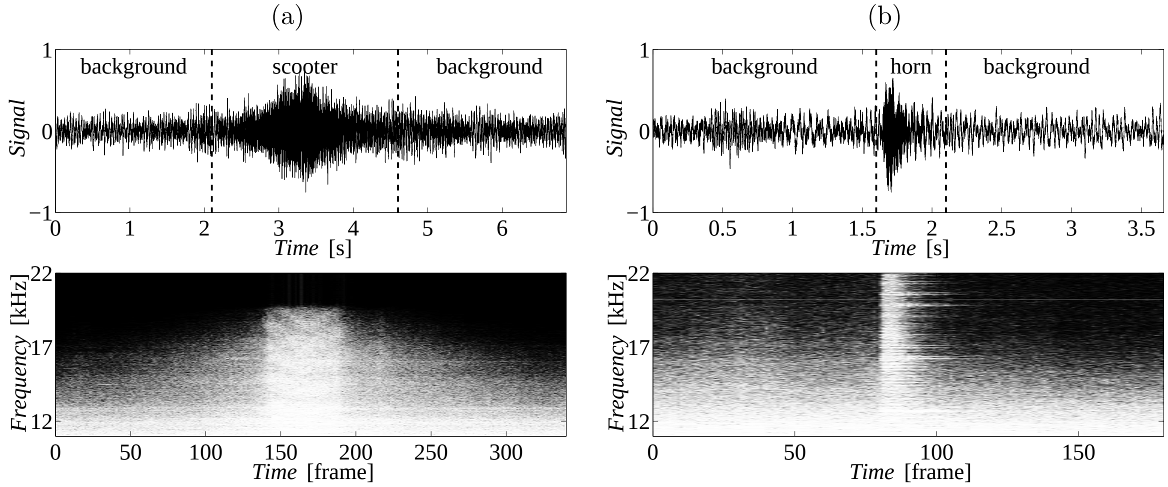

Figure 2: *Time signal (upper plots) and spectrogram (lower plots) for (a) a scooter pass-by and (b) a horn signal.*

decorrelate the coefficients. There exist various techniques to reach these goals. In this work, we adopt the Perceptual Linear Prediction (PLP) algorithm [3], which was originally developed for analyzing speech signals.

The following step consists in computing an auditory spectrum by integrating the power spectrum within perceptually meaningful bandpass filters. That is, every auditory spectral coefficient is obtained by weighting and summing the power spectral coefficients of the corresponding frequency band. In this work, we adopt trapezoidally shaped and overlapping filters as defined in [3]. The number of bands is set to 15.

Although the next steps are optional, they have shown to improve the performance. An equal-loudness curve is applied to the auditory spectrum. It consists in highpass filtering the auditory spectrum in order to approximate the unequal sensitivity of the human hearing. An important effect is to eliminate the contribution of a DC offset in the time signal. Besides, the range of the auditory spectrum coefficients is compressed by taking their cubic root.

The last but one step consists in computing cepstral coefficients from the auditory spectral coefficients. This step allows reducing again the feature variability and obtaining uncorrelated coefficients. They are computed as follows. First, autocorrelation coefficients are computed by taking the inverse FFT of the auditory spectrum. Next, a Linear Prediction Coding (LPC) model [2] is efficiently derived from the autocorrelation coefficients by solving the Yule-Walker equations with the Durbin-Levinson algorithm. In this work, the order of the LPC model is set to 10. Finally, the LPC coefficients are orthogonalized by converting them to cepstral coefficients using standard recursions. In this work, the number of cepstral coefficients is set to 13.

The final step consists in estimating time derivatives of the cepstral coefficients, the so-called delta cepstral coefficients [4]. They tend to emphasize the dynamic aspects of the power spectrum over time and are typically used as an add-on to the static cepstral coefficients. First-order time derivatives are computed by locally correlating each cepstral coefficient sequence with a straight line, while second-order time derivatives correspond to a similar correlation but with a parabolic function.

We ultimately represent the time signal as a sequence of feature vectors that contain cepstral coefficients augmented with delta cepstral coefficients. The first cepstral coefficient is not used in order to make the ANR system insensitive to the recording level while its first- and second-order time derivatives are included. The other cepstral coefficients as well as their first-order time derivatives are used. These vectors are next fed to a classifier whose task is to locally identify the nature of the noise source as described in section 2.3.

Spectral Enhancement: Wiener filter

We can model the time signal $x(n)$ provided by the audio interface as the sum of a noise event signal $v(n)$ and a background noise signal $w(n)$. Under this assumption, we can write that

$$P_x(k,l) = P_v(k,l) + P_w(k,l) \tag{1}$$

where $P_x(k,l)$, $P_v(k,l)$ and $P_w(k,l)$ denote the power spectrum of $x(n)$, $v(n)$ and $w(n)$, respectively, for the l-th frame and the k-th discrete frequency. The energy ratio between the noise event signal and the background noise signal can be sometimes very low. It results that the power spectrum of the noise event signal can be "swamped" in the power spectrum of the background noise signal. This makes the detection and the identification of noise events more sensitive. In this work, we suggest to apply a spectral enhancement algorithm in order to improve the recognition performance. This algorithm can be viewed as an optional step which takes place between the second step and the third step. It consists in weighting $P_x(k,l)$ such that the regions where the noise event signal is dominant are enhanced while the others are attenuated. Such processing can be implemented by a generalized Wiener filter [2] and is computed for every frame and every frequency as follows:

$$H_{\text{Wiener}}(k,l) = \max\left(\frac{P_x(k,l) - \alpha\widehat{P}_w(k,l)}{P_x(k,l)}, \beta\right)^{\gamma} \tag{2}$$

where the parameters α, β and γ stand for the overestimation factor, the spectral floor factor and the gain exponent, respectively. In this work, we set $\alpha = 1.25$, $\beta = 0.01$ and $\gamma = 2.5$. The estimate $\widehat{P}_w(k,l)$ of the power spectrum of the background noise signal is adaptively computed when only background noise signal is present, that is,

$$\widehat{P}_w(k,l) = \begin{cases} \delta\widehat{P}_w(k,l-1) + (1-\delta)P_x(k,l) & \text{if } \displaystyle\sum_k P_x(k,l) \leq \eta \sum_k \widehat{P}_w(k,l-1), \\ \\ \widehat{P}_w(k,l-1) & \text{if } \displaystyle\sum_k P_x(k,l) > \eta \sum_k \widehat{P}_w(k,l-1). \end{cases} \tag{3}$$

The procedure is initialized with $\widehat{P}_w(k,0) = P_x(k,0)$, $\forall k$. In this work, we set the smoothing parameter $\delta = 0.9$ and the threshold parameter $\eta = 2.0$. Once $P_w(k,l)$ has been estimated, the filter $H_{\text{Wiener}}(k,l)$ can be derived. The enhanced power spectrum $\widehat{P}_v(k,l)$ is eventually obtained by applying the filter to the observed power spectrum $P_x(k,l)$,

$$\widehat{P}_v(k,l) = H_{\text{Wiener}}(k,l)P_x(k,l). \tag{4}$$

In figure 3.(a), we show the power spectrogram of the a horn signal (as already presented in figure 2) and its enhanced version obtained by Wiener filtering. Clearly, the power spectrum of the noise event signal is "dug up": the areas in the time-frequency plan corresponding to the noise event are accentuated.

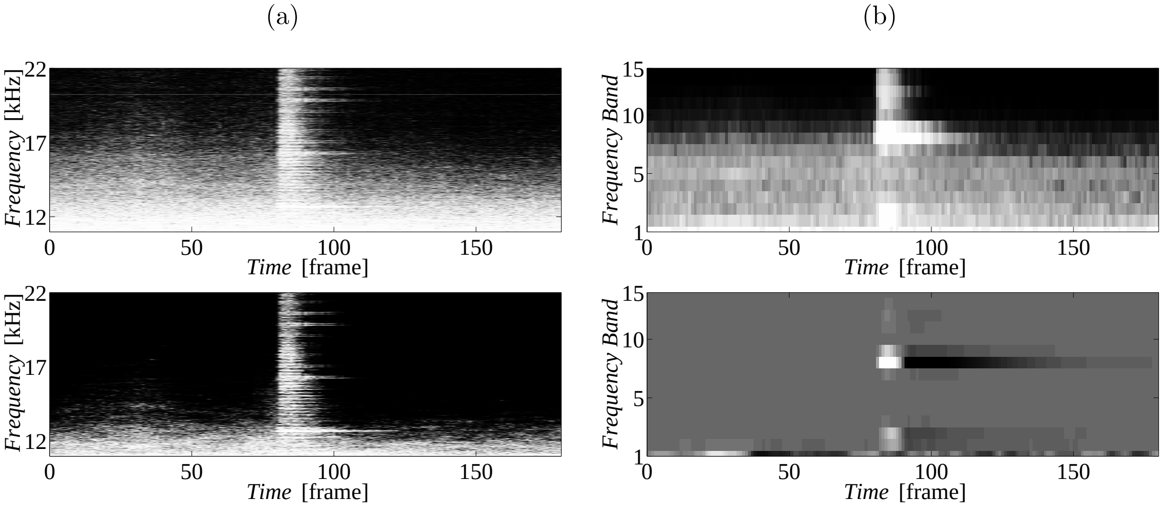

Figure 3: *(a) Power spectrogram of a horn signal and its enhanced version by Wiener filtering; (b) auditory spectrogram of a horn signal and its enhanced version by J-Rasta filtering.*

Spectral Enhancement: J-Rasta filter

Define $A_x(b,l)$ as the auditory spectrum of the time signal $x(n)$ at the l-th frame and the b-th frequency band. Like for the power spectrum, we can write

$$A_x(b,l) = A_v(b,l) + A_w(b,l) \tag{5}$$

where $A_v(b,l)$ et $A_w(b,l)$ stand for the auditory spectrum of $v(n)$ and $w(n)$, respectively. In our recordings, we observe that the auditory spectrum is stationary most of the time and significant variations occur when noise events happen. In other words, the background noise signal contributes to the observed auditory spectrum as constant or slowly varying components while the noise event signals cause more rapidly varying components. This suggests that the background noise contribution can be reduced by filtering the time sequence of every auditory spectrum coefficient. Such approach was originally proposed to enhance speech signals in noise. It is known as a J-Rasta filter [5] and can be viewed as an optional step which takes place between the third step and the fourth step. The J-Rasta filter is implemented as follows. First, a nonlinearity mapping is applied to the auditory spectrum coefficients,

$$A_{x,\mathrm{NL}}(b,l) = \log\big(1 + J \times A_x(b,l)\big). \tag{6}$$

Then, the time sequence of every mapped auditory spectrum coefficient is linearly convolved with a bandpass filter $h_{\mathrm{Rasta}}(l)$ and the nonlinearity mapping is inverted,

$$A_{x,\mathrm{J-Rasta}}(b,l) = \exp\big(A_{x,\mathrm{NL}}(b,l) *_l h_{\mathrm{Rasta}}(l)\big). \tag{7}$$

In this work, we use the bandpass filter defined in [5], which emphasizes frame-to-frame spectral changes between 1 Hz and 10 Hz. We set the constant $J = 1 \times 10^{-6}$.

In figure 3.(b), we show the auditory spectrogram of a horn signal and its enhanced version obtained by J-Rasta filtering. We observe that the energy areas corresponding to background noise are reduced while rapid variations related to the noise event are more strongly marked.

2.3 Local Classification

Any signal recorded in a urban environment can be viewed as a sequence of acoustic states which can correspond either to noise events or to background noise. For a given recording,

Table 1: *Distribution of noise recordings into the training set and the test set.*

Set	Scooter	Horn	Scooter + Horn
Train	338	393	731
Test	86	100	186

the number of states and their duration are unknown. In order to recognize noise events, we need to find their limits (detection) and their nature (identification). The primary step consists in assigning every feature vector to a state. To do this, we need to estimate a model of the statistical distribution of the feature vectors for every state. Various approaches are possible, *e.g.*, codebook-based non-parametric models [6] or Gaussian Mixture Models (GMMs) [7]. In this work, we adopt an approach based on an Artificial Neural Network (ANN). More exactly, a Multi Layer Perceptron (MLP) [4] is used. When presented with a feature vector, the MLP estimates the *a posteriori* probability $P(\text{state}|\text{vector})$ for every state. The highest probability gives the state of the feature vector.

The MLP coefficients can be estimated by supervised training. That is, the MLP is presented with feature vectors for which the correct state is known. Then, the MLP coefficients are adapted in order to minimize the squared error between the actual outputs and the expected *a posteriori* probabilities, namely 1 for the correct state and 0 for the others.

The classification performance can be improved by feeding the MLP not only with the current feature vector to be classified but also left and right context vectors. In this work, we consider a symmetric context and the number of context vectors is set equal to 9. Other important parameters are the number of hidden layers in the MLP and their size. In this work, we use a single hidden layer and its size is set equal to 200.

2.4 Global Decoding

The local classification module searches the most likely state for every feature vector independently. These decisions are likely to change inconsistently, even inside a noise event. The global decoding module aims at searching the most likely sequence of states globally given the sequence of observed feature vectors and the corresponding lattice of state probability scores. To do so, it relies on Hidden Markov Models (HMM) where the states are assumed to be linked together forming a Markov chain [2]. Such approach has already been used for ANR with discrete HMMs based on codebooks [6] and with continuous HMMs based on GMMs [7]. We naturally extend it to hybrid HMMs based on MLPs being inspired by what has been done in ASR [4].

3 EXPERIMENTAL RESULTS

The audio material that is used in the experiments consists of 917 recordings in urban environments. Every recording contains background noise with a single noise event, either a scooter by-pass or a horn signal, in its middle. The recordings were gathered thanks to the 01dB acquisition hardware *Symphonie* powered the 01dB software *dBTrig32* in Paris and Lyon downtowns. They were separated into a training set and a test set as detailed in table 1. Every recording was segmented by hand, that is, the time limits of its noise event were manually identified.

Various feature extractions were applied to the training set and MLP's were trained in a supervised mode using the corresponding segmentations. The same feature extractions were applied to the test set and the corresponding trained MLP's were used to recognize it. The recognition errors are counts as follows. Assume that N noise events have been detected for a

Table 2: *Recognition results (C: correctly recognized events / S: substitutions / D: deletions / I: insertions) for various feature extractions on (a) the training set and (b) the test set. The figures between squared brackets correspond to the recognition and error rates in percent.*

(a)

Feature Extraction	Scooter		Klaxon	
	C/S/D/I	[%]	C/S/D/I	[%]
PLP	336/3/9/9	[99.4/0.9/2.7/2.7]	346/21/26/26	[88.0/5.3/6.6/6.6]
PLP + Wiener	333/2/3/3	[98.5/0.6/0.9/0.9]	368/21/34/34	[93.6/5.3/8.6/8.6]
PLP + J-Rasta	334/3/1/1	[98.8/0.9/0.3/0.3]	377/7/9/9	[95.9/1.8/2.3/2.3]
PLP + Wiener + J-Rasta	336/2/0/0	[99.4/0.6/0.0/0.0]	375/11/7/7	[95.4/2.8/1.8/1.8]

(b)

Feature Extraction	Scooter		Klaxon	
	C/S/D/I	[%]	C/S/D/I	[%]
PLP	79/3/4/4	[91.9/3.5/4.6/4.6]	88/7/5/5	[88.0/7.0/5.0/5.0]
PLP + Wiener	79/5/2/2	[91.9/5.8/2.3/2.3]	85/11/4/4	[85.0/11.0/4.0/4.0]
PLP + J-Rasta	82/4/2/2	[95.3/4.6/2.3/2.3]	93/5/2/2	[93.0/5.0/2.0/2.0]
PLP + Wiener + J-Rasta	81/5/0/0	[94.2/5.8/0.0/0.0]	94/6/0/0	[94.0/6.0/0.0/0.0]

given recording. If $N = 0$, a deletion is counted. If $N \neq 0$, we search for the detected noise event that overlaps the most the actual noise event. If we do not find any detected event, N insertions and a deletion are counted. Otherwise, we compare the actual event and the most overlapping detected event. If they do not match, then a substitution is counted, else the recognition is correct. All the results are given in table 2 for both the training set and the test set, and separately for each type of noise. For each recognition experiment, we report the numbers of correctly recognized events, substitutions, deletions and insertions, as well as the recognition and error rates obtained by dividing the previous numbers by the number of events.

We observe that the proposed system performs satisfyingly with many correctly recognized noise events and few substitutions. Car horns and scooter events are both identified with a high success rate, which ranges from 85% to 95% depending on the feature extraction. Also, this rate is significantly improved by using a spectral enhancement filter. The J-Rasta filter appears to yield better results than the Wiener filter in these experiments. Combining both filters does not seem to provide further improvement.

It is also interesting to look at the detection errors, *i.e.* the errors $\Delta_{\text{start}} = \hat{t}_{\text{start}} - t_{\text{start}}$ and $\Delta_{\text{stop}} = \hat{t}_{\text{stop}} - t_{\text{stop}}$ between the estimated boundaries $(\hat{t}_{\text{start}}, \hat{t}_{\text{stop}})$ of the detected event as provided by the recognition process and the reference instants $(t_{\text{start}}, t_{\text{stop}})$ as known from the handmade segmentation. To do this, we plot in figure 4 the empirical cumulative distribution functions of the detection errors for correctly recognized noise events.

We see that car horns are detected with a higher precision than scooter events are. A car horn is indeed a better time-limited noise event, while the starting and ending instants of a scooter event are more difficult to assess, even by the human ear.

4 CONCLUSIONS

In this communication, we presented a system for Automatic Noise Recognition (ANR). It first extracts feature vectors from the time signal that represent the time evolution of the power spectrum in a compact form. This feature extraction can be completed by spectral enhancement procedures that emphasize the representation of the noise events with respect to the background noise. The feature vectors are then presented to an Artificial Neural Network

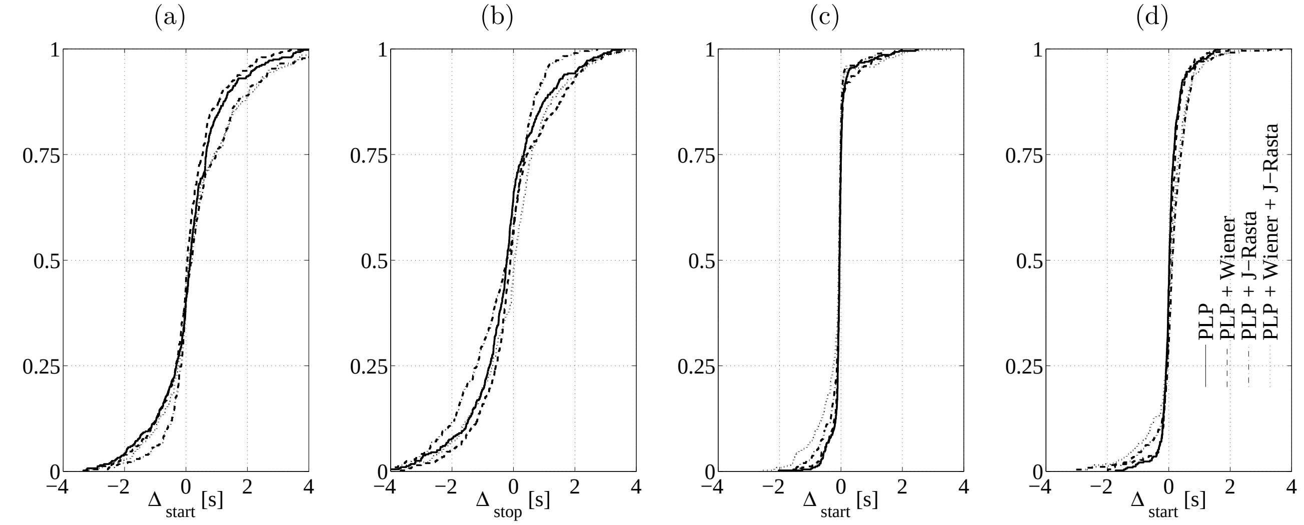

Figure 4: *Empirical cumulative distribution function of the detection errors* Δ_{start} *and* Δ_{stop} *for (a)-(b) scooter events and (c)-(d) horn events.*

(ANN) that classifies each of them with respect to the nature of the noise source. These local decisions are finally integrated in the framework of Hidden Markov Models (HMM) to get a global decision about the localization and the nature of the noise events within the recordings. The proposed system was tested for recognizing scooter pass-bys and horn signals recorded in urban environments. High identification rate and satisfying detection accuracy were reported. We also showed that spectral enhancement significantly improved the performance.

Research directions for future work include the test of the proposed ANR system to more difficult task (more noise types) and the use of confidence scores for rejecting mis-recognized (badly detected and/or wrongly identified) noise events. Besides, the algorithm will be integrated on a DSP so as to incorporate it in urban monitoring systems.

REFERENCES

[1] “Directives 2002/49/EC Relating to the Assessment and Management of Environmental Noise”, *Official Journal of the European Communities*, vol. L 189, pp. 12–25, Jul. 18, 2002.

[2] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, 2nd ed., 2000.

[3] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech”, *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[5] H. Hermansky and N. Morgan, “RASTA Processing of Speech”, *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[6] P. Gaunard, C. G. Mubikangiye, C. Couvreur and V. Fontaine, “Automatic Classification of Environmental Noise Events by Hidden Markov Models”, *Proc. of ICASSP*, vol. 6, pp. 3609–3612, Seattle, USA, May 1998.

[7] M. Casey, “MPEG-7 Sound Recognition”, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, Jun. 2002.