The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Underwater Transient and Non Transient Signals Classification Using Predictive Neural Networks

Yan Guo and Bruno Gas

Abstract— **The project ASAROME (Autonomous SAiling Robot for Oceanographic MEasurements) is working on a small autonomous sailboat in order to make measurements and observations in the marine environment for long periods. In this project, perception plays an important role by giving an estimate of the speed of surface winds, the state of the sea surface and the rate of precipitation in wet weather. In this paper, the unknown signals are first encoded with different codes (ERB, MFCC, LPC, LPCC). Then the coded signals are modeled by two different methods of classification: predictive and k-Nearest Neighbor. The final part of the system uses local and global decision to recognize the class of the unknown signal. Experiments are conducted to compare the results obtained by different encodings. Our results show that MFCC does not represent the ideal approach for the recognition of underwater audio signals, but LPCC seems to be a better candidate.**

I. INTRODUCTION

ASAROME (Autonomous SAiling Robot for Oceanographic MEasurements) is a research project focused on autonomous robotics. The project aims to prove the relevance of using sailing autonomous surface vehicles (ASV) for long (several weeks) observation and measurement missions in marine environments. Based on a robotized sailing boat concept from Robosoft, the ASAROME project focuses on adding and integrating advanced functionalities in the fields of aero and hydrodynamics modeling, as well as action/perception in robotics, to build a sailed autonomous surface vehicle demonstrator.

One of the tasks of the project is the multiperception coupling task which gathers the following detection methods: panoramic vision, radar, inertial and gyro sensors. It will be used for detecting obstacles (boats, drifting floating bodies) and for estimating the sea state (wave direction and amplitude).

In the field of perception, most of the literature to date relating to the detection of obstacles at sea concerns the problem of tracking and monitoring of appropriate paths in order to avoid collision situations. The anticollision maneuvers are mainly based on the route of radar echoes observed on moving objects. The design of the ARPA system [1], initiated in the early 80s, had the primary purpose of the automation of obstacle monitoring and planning safe trajectories. It continues today with the introduction of artificial intelligence tools [2]. For example, we found with the Syllogic sailing lab [3] the implementation of predictive algorithms to predict the relative height and direction of nearby waves, from the fusion of data with the sensors placed in the boat (a measure of the strength and direction of winds, accelerometers, etc.). These studies work on correlating the data related to the state of the sea and wind with data as detected on the boat. They do not use visual and/or audio sensor data.

We are specifically interested in the data resulting from underwater sound sensors with the objective to detect near and far motor vehicles. In this context, we propose a comparative study of coding and classification algorithms commonly used in the audio field for the classification of underwater sound events (noise related to weather conditions, the maritime traffic or the proximity of marine animals, etc.). Lim et al. [4], [5], have recently shown that it was possible to classify underwater transient sound events by the Mel Frequency Cepstral Coefficients (MFCC) features of acoustic frames. They proposed a classification of feature vectors by comparing Euclidean distances (k-NN), or by learning of a Multilayer Perceptron (MLP). We propose in this article to extend the work of Lim et al. in the following way:

- extension to the case of signals *non transient* or *long-term*, i.e. whose characteristics vary slightly during time.
- study of other coding methods (Linear predictive coding (LPC), Linear prediction cepstral coefficents (LPCC) and Equivalent Rectangular Bandwidth (ERB));
- classification by non-linear predictive modelling;
- more signal classes (13 to 30 instead of 8 are used in [4], [5]).

In the first part of our paper, we describe the signal coding algorithms. In the second part, we discuss the classification of underwater signals. The third section of this article contains the description of experiments and the analysis of the results obtained.

II. THE COMPOSITION OF THE SYSTEM

A. Representation of signals

For the classification problem covered in this paper, we suppose that the shape, duration and spectral response of the signal are not known. The spectral features may be highly variable in time, which has led many authors to use time-frequency representation (wavelet transform [6], Wigner-Ville distribution and Cross Wigner-Ville distribution [7], [8], or short-time Fourier transform (STFT)) methods.

Tucker and Brown [9] proposed another idea to classify underwater transient signals recorded by passive sonar. They propose to consider perceptual acoustic features, i.e. those which contain information that human listeners are likely to use in transient classification tasks.

UPMC Univ Paris 06, UMR 7222, F-75005, Paris, France, 4 Place Jussieu, BP 173, 75252 Paris Cedex 05, FRANCE. This work is funded under the project ANR ASAROME (Num. ANR-07-ROBO-0009)
Guo@isir.fr, Bruno.Gas@upmc.fr

In this paper, we follow the approach of Lim et al [4], [5], who use conventional methods of coding (MFCC) and the classifiers of k-NN and MLP to classify the underwater transient signals. For that, we tested four types of codes: three of which are commonly used in audio signal processing (MFCC, LPC and LPCC) and an approach based on the model of the cochlea developed by Patterson [10] (ERB).

All methods used are based on a short-term description of the spectrum, the signal is divided into successive frames of about 45 ms. A pretreatment by using Hamming Windows is then carried out on each frame, before the extraction of 14 parameters according to the 4 previous encoding methods.

1) ERB: The ERB filter is a general purpose model of peripheral auditory processing that produces *auditory images* of sounds. The model includes the interactions associated with adaptation and suppression as observed in the auditory nerve, and also the phase alignment and temporal integration which takes place before the formation of initial sound images, but does not include any of the processes that combine information across widely separated frequency bands. Each filter models the signal present at the output of a nerve of the cochlea. For N ERB filters whose frequency responses range as shown in Figure 1, we obtain N signals for each frame, which we use to calculate the energy during a period of time corresponding to the frame.

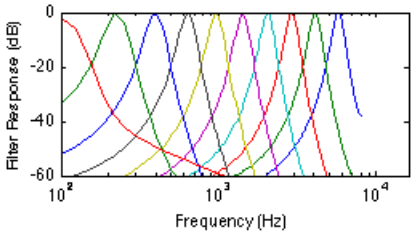

Fig. 1. Impulse responses of a set of filters performing ERB on audio signals.

2) LPC, LPCC and MFCC: These three feature extraction methods are classically used in the literature for compression or classification of audio signals, especially speech signals. LPC is founded on modelling the vocal tract, following the hypothesis of linear source-filter. LPCC calculate the coefficients of the cepstral representation of the signal from the LPC coefficients. Finally, MFCC extracts the feature frequency of signals following a non-linear scale called *Mel scale* and inspired by the auditory sensitivity of the human ear. We are interested in this appoach because it is commonly used in speech recognition. In particular, Lim et al. [5] use this for classification of transient signals, but we show that it is possibly not the most effective method.

A question arises which is to know if MFCC and LPC are suitable coding in our application given the fact that we consider both speech and non-speech signals. As an answer one can say that MFCCs are suitable for speech and non-speech signals since they are inspired by the human ear. On the other hand LPCs could be less appropriate since they are designed to model the vocal tract. But unlike MFCCs they are better adapted to underwater sounds for which frequency bandwidth are somewhat different. This is the reason why we suggest in this article to use both MFCC and LPC coding methods.

B. Classification of signals

We propose two classification methods for audio signals. This allow us to both compare the techniques, and also provide more robust information from the environment on the decision and then command steps of the complete robotic system. The first classification technique uses the predictive modelling of signals. This method is commonly used in speaker recognition systems [11], which has the advantage of making a single decision from a number of considered frames. The second method of k-NN is one of the simplest appoaches classically used in pattern recognition. But it requires the set up of a global decision making algorithm, that uses the local decisions obtained at the level of the acoustic frame.

1) Predictive modelling: Predictive modelling of sound sources allows us to estimate the distance between an unknown source and a set of models of sources. These sources are each modeled by a MLP network, and used for prediction. Consider the vector of parameters $\mathbf{x}_k$ extracts a frame k. The modelling of a sound source trains the network by minimizing the square error of prediction that is calculated on all frames of the sequence:

$$Q(\Omega) = \sum_k \|F(\mathbf{x}_{k-2}, \mathbf{x}_{k-1}) - \mathbf{x}_k\|^2 \qquad (1)$$

where F represents the function performed by the network, with two successives frames $\mathbf{x}_{k-2}$ and $\mathbf{x}_{k-1}$, which are associated with $\mathbf{x}_k$, are used as input.

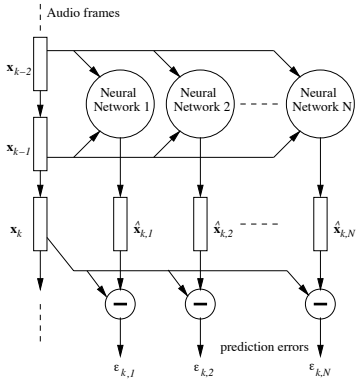

Fig. 2. Predictive architecture with neural networks.

For the problem of classification of a source among N, we have N networks previously trained with signals belonging to the same sound class (Figure 2). In recognition, an unknown source with its frames presented as input to the N networks. N prediction errors are calculated and the final decision is obtained according to the principle of maximum likelihood: the class of the unknown source is the network with the smallest prediction error.

2) Classification k-NN: The k-NN classifier allows us to carry out a local classification of acoustic frames: for a given frame, the feature vector is compared with all the vectors of a referenced base which was established previously. Euclidean distance between the unknown and referenced encoded signal is calculated. The decision is reached by assigning the encoded signal according to the majority classification of the k closest referenced vectors. For a frame sequence, we arrange a series of local decisions from which a global decision is made, as used by Lim et al. [4].

3) Local and global decisions: The two classification algorithms (the predictive modelling and k-NN) produce local decisions at frame level. This local decision of the k^{th} acoustic frame is represented by c_k. For the predictive classifier, we have :

$$c_k = \arg \min_{i=1,M} \{\epsilon_{k,i}\} \tag{2}$$

where $\epsilon_{k,i}$ designates the N errors calculated from N multilayer perceptrons. The global decision is obtained by:

$$c = \arg \min_{i=1,M} \{\sum_{k=1}^{K} \epsilon_{k,i}\} \tag{3}$$

where K represents the total number of frames of the unknown audio sequence. We will designate this decision algorithm by $GD1$.

The k-NN appoach does not allow us to compute a global decision so easily. A *majority decision* algorithm is often used in the literature:

$$c = \arg \max_{i=1,M} \{|U_i|\} \tag{4}$$

where U_i is the set of all the frames belonging to class i, $i = 1, \ldots, N$. We will designate this decision algorithm by $GD2$.

The last algorithm, $GD3$, is variant of $GD2$ consisting of considering the longest chain of successive identical local decisions.

III. EXPERIMENTATIONS

We present a set of simulations taken from a database of real recorded signals, *The Underwater Sound Effects Series* [12]. This database includes 500 sound effects which were recorded with a sampling frequency of 44KHz from an underwater perspective using a pair of Brüel & Kjaer hydrophones. Among these recordings, some correspond to transient sound events (of a maximum duration in an order of seconds) and others to non transient (up to one minute of recording). Two bases have been established from this album, one composed of transient signals (13 classes of signals) and the other composed of non transient signals (30 classes of signals). Every class includes 4 signals datasets for training and 2 signals for test. This database allows us to consider both types of events. Generally, the transient signals are often covered in the literature. But in the robotic context that concerns us, we have to deal with non transient signals also, for example to provide sea state or ship information. Many of these signals were recorded both in water and above water. Figures 3 and 4 represent respectively the spectrum of signals obtained by recording an event of non transient type *drip*, and an event of transient type *object hitting metal*. The top graph of each figure represents the spectrum of registration in the water and the bottom one is the spectrum in the air.

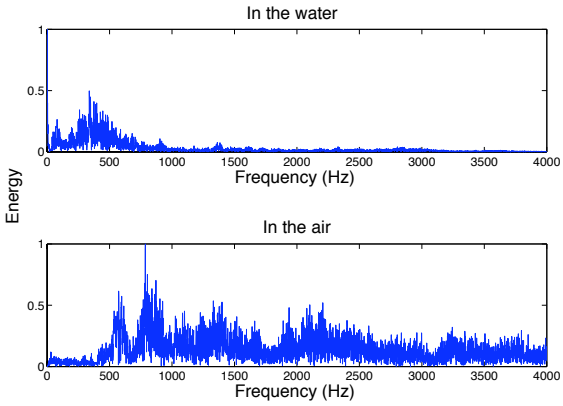

Fig. 3. Spectrum of a sound event corresponding to a *drip recorded in the water (above) and in the air (below).*

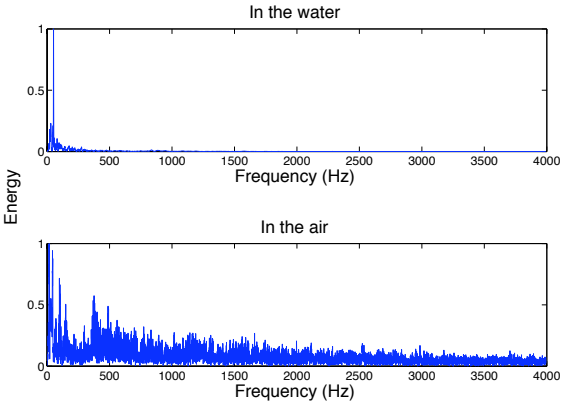

Fig. 4. Spectrum of a sound event corresponding to a hit of the metal recorded in the water (above) and in the air (below).

We can see the spectral differences of these figures, essentially a reduction in the bandwidth of signals in the water, and a shift towards low frequencies. From this we know that lower frequencies propagate better than high frequencies under water.

From this, we can conclude that the MFCC method, classically used in speech recognition, is probably less well adapted to the analysis of underwater signals. This is verified in the experiments presented in the next section.

A. Encoding stage

For the four methods used, we represented the evolution of coefficients throughout the duration of an event. In all

simulations, we used 14 coefficients. We have studied the optimum dimensionality of the encoded signals. We found that the optimum dimensionality varies between 10 and 18 with an average of 14, depending on the nature of the signal (transient or non transient), and the feature extraction method used. Figures 5 and 6 show examples of evolution for a non transient and a transient signal respectively.

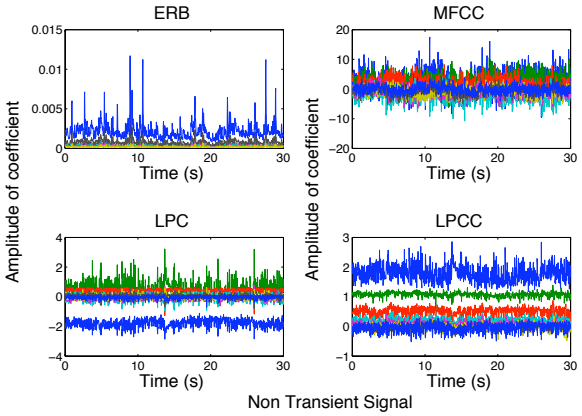

Fig. 5. Evolution of coefficients ERB, LPC, LPCC and MFCC during time for a non transient signal(30s, windows of 2048 sample points).

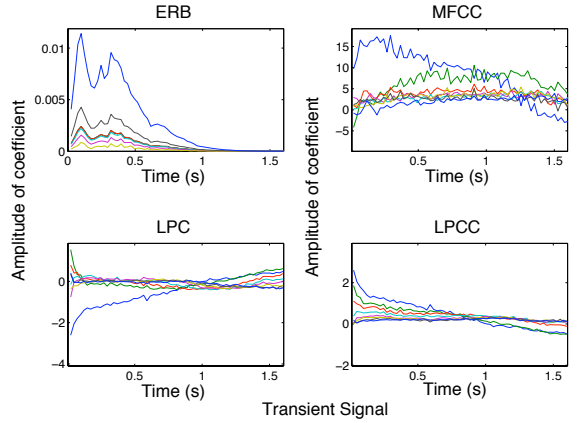

Fig. 6. Evolution of coefficients ERB, LPC, LPCC and MFCC during time for a transient signal(1.6s, windows de 2048 sample points).

In the case of non transient signals, there is a more important variance of MFCC coefficients than found in LPC and LPCC coefficients. This shows the inadequacy of MFCC coefficients. In fact, throughout the evolution of LPC and LPCC coefficients, the properties of the non transient signal are quasi-stationary. In contrast, a large part of the MFCC frequency bands (frequencies above 1500Hz as shown in Figure 3) does not convey useful information. In the same way, the ERB code covers the entire spectrum of audible frequencies. Many of the coefficients have low amplitudes because the energy of higher frequency channels is low.

With the transient signal, Figure 4 shows that they are narrow-band signals. The evolution of coefficients is presented in Figure 6. They show that, the transient signals are not stationary. The number of representative evolution coefficients is more important in the case of LPC and LPCC than in MFCC and ERB. This shows that LPC and LPCC coefficients may be better candidates for the modelling of underwater signals. These individual results were consistently observed throughout the majority of signals available in the database.

B. Classifier parameters

The classifiers proposed in this paper (the predictive neural and the k-NN) need to be parametrized correctly. For this reason, we realized a series of simulations with the number of hidden cells of neural networks on the one hand, and with the number of neighbors to k-NN classifier on the other. Two bases have been established, one composed of transient signals (13 classes of signals) and the other composed of non transient signals (30 classes of signals). We have conducted two sets of experiments. The separation of bases was necessary because the transient signals and non transient have different features which may justify the use of different techniques. We present here two simulation examples: Figure 7 represents the scores obtained with the predictive classifier and the LPCC codes according to the number of hidden cells, and Figure 8 represents the scores obtained with the k-NN classifier and the MFCC approach according to the number of neighbors. For these tests, the decision algorithms $GD1$, $GD2$ and $GD3$ have been tested with the predictive classifier and $GD2$, $GD3$ with the k-NN classifier. This is because the k–NN that we have implemented does not allow the use of $GD1$ (see earlier in this article).

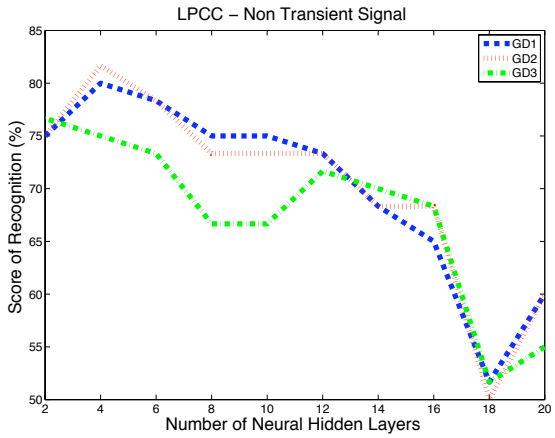

Fig. 7. Recognition rate by the number of hidden cells using LPCC in the base of non transient signals.

In general, it should be noted that a small number of neurons or a small number of neighbors make it possible to obtain the best scores. Concerning predictive networks, the problem of over-learning, when the number of cells is too large, the networks learn the training data well, but generalize badly. Concerning the k-NN classifier, this is due to the low number of references available in the base of signals. However we know that k-NN works better than when the number of references is large.

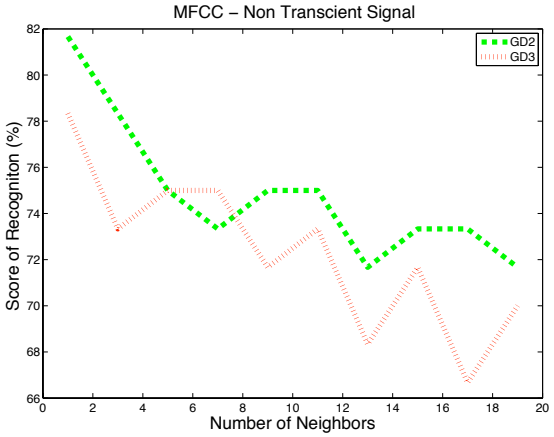

Fig. 8. Recognition rate by the number of hidden cells using the codes MFCC with the base of non transient signals.

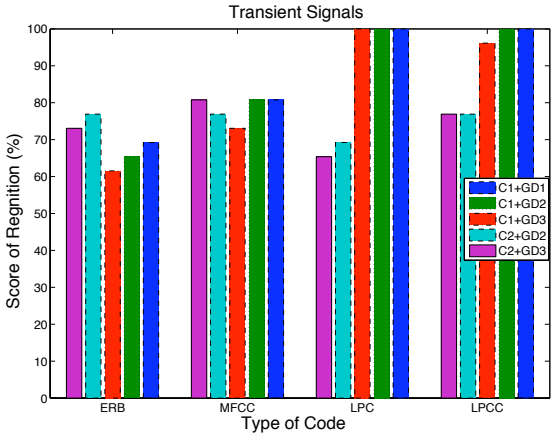

Fig. 10. Recognition rate with the codes of the transient base.

C. Experimental results

The last set of experiments in this article concern the scores obtained by the two classifiers tested using four ERB, LPC, LPCC and MFCC with the two bases of transient and non transient signals. Concerning the predictive networks, scores are obtained by using a test data set not used in the training process. Figures 9 and 10 represent the scores obtained. $C1$ is the predictive classifier and $C2$ represents the k-NN classifier.

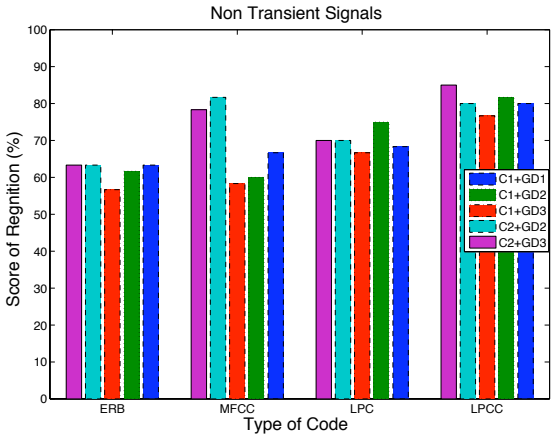

Fig. 9. Recognition rate of the codes with the base of non transient signals.

With the non-transient signals, LPCC returns superior results regardless of the classification algorithms used (between 78% and 85%). In contrast, results obtained with MFCC are more mixed, good with the k-NN classifier (between 79% and 81%), but poorer with the predictive classifier (between 59% and 65%).

For the transient signals, the difference vis a vis of classifier used increases. LPC and LPCC always obtains the best results, when using predictive networks. The predictive network result makes it possible to model the evolution of non-stationary signals and this is interesting because the transient signals are non-stationary. Finally, the obtained score of 100%, which is higher than the score obtained with non transient signals, is explained by the number of classes: only 13 for the transient signals and 30 for the non transient signals. Unfortunately these class numbers are too small to say that LPC and LPCC are prefect for this task.

These results show that MFCC does not represent the ideal approach for the recognition of underwater audio signals. LPCC seems to be a better candidate. This result can be explained by the fact that MFCC code is only suitable for the spectrum of the speech signals in the air. For example, Figure 3 shows that the spectral ranges are different in the air and in the water so that MFCC appears to be not well adapted. In contrast, temporal codings like LPCs are adapted to the spectral envelope of the signals. The continuation of our work within project ASAROME consists of the fusion of multimodal data (audio, visual and accelerometers, etc.) for aiding decision-making and the development of suitable commends for the navigation of the autonomous boat.

IV. CONCLUSIONS AND PERSPECTIVES

In this paper, we presented algorithms for classification of audio signals to assist in the navigation of an autonomous boat. We compared two classifiers and four methods of signals coding, expanding upon work started by Lim et al. in [4], [5]. Results show that if the methods of signal representation are well chosen, it is possible to obtain promising results in classification. In particular, the MFCC method typically used in speech processing are not suitable for underwater audio signal processing.

REFERENCES

[1] "Imo preference standards for automatic radar plotting aids (arpa)," *Resolution A. 422 (XI)*, 1979.

[2] R. Smierzchalski and Z. Michalewicz, "Adaptive modeling of a ship trajectory in collision situations at sea," in *Evolutionary Computation Proceedings, IEEE Congress on Computational Intelligence*, 1998, pp. 342–347.

[3] M. van Aartrijk and C. Tagliola, "Ai on the ocean: the robosail project," in *Proc. of European Conference on Artificial Intelligence*, 2002, pp. 653–657.

[4] T. Lim, K. Bae, C. Hwang, and H. Lee, "Classification of underwater transient signals using mfcc feature vector," pp. 1– 4, 2007.
[5] ——, "Underwater transient signal classification using binary pattern image of mfcc and neural network," vol. E91-A, no. 3, pp. 772– 774, 2008.
[6] D. M. S. and W. J., "Improved transient signal detection using a wavepacket-based detector with an extended translation-invariant wavelet transform," in *IEEE trans. signal process*, vol. 45, 1997, pp. 841–850.
[7] N. Yen, "Time and frequency representation of acoustic signals by means of the wigner distribution: Implementation and interpretation," vol. 81, no. 6, pp. 1841 – 1850, 1998.
[8] B. Boashash and P. Shea, "A methodology for detection and classification of some underwater aoustic signals using time-frequency analysis techniques," vol. 38, no. 11, pp. 1829 – 1841, 1990.
[9] S. Tucker and G. Brown, "Classification of transient sonar sounds using perceptually motivated features," vol. 30, no. 3, pp. 588 – 600, 2005.
[10] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C.Zhang, and M. Allerhand, "Complex sounds and auditory images," pp. 429–446, 1992.
[11] A. Melouk and P. allinary, "A discriminative neural predictive system for speech recognition," in *ICASSP*, vol. 2, 1993, pp. 533–536.
[12] (2004) Underwater sound effet series. [Online]. Available: http://www.sound-ideas.com/underwater.html