Proceedings of Australasian Conference on Robotics and Automation, 2-4 Dec 2013, University of New South Wales, Sydney Australia

Teaching a robot to hear: a real-time on-board sound classification system for a humanoid robot

Thomas D'Arcy, Christopher Stanton, and **Anton Bogdanovych**
MARCS Institute, University of Western Sydney
16107410@student.uws.edu.au, c.stanton@uws.edu.au, a.bogdanovych@uws.edu.au

Abstract

We present an approach for detecting, classifying and recognising novel non-verbal sounds on an Aldebaran Nao humanoid robot. Our method allows the robot to detect novel sounds, classify these sounds, and then recognise future instances. To learn the names of sounds, and whether each sound is relevant to the robot, a natural speech-based interaction occurs between the robot and a human partner in which the robot seeks advice when a novel sound is heard. We test and demonstrate our system via an interactive human-robot game in which a person interacting with the robot can teach the robot via speech the names of novel sounds, and then test the robot's auditory classification and recognition capabilities by providing further examples of both novel sounds and sounds heard previously by the robot. The implementation details of our acoustic sound recognition system are presented, together with empirical results describing the system's level of performance.

1 Introduction

For most people, hearing plays an important role in everyday life. While hearing is crucial for understanding speech, non-verbal sounds are also a valuable source of perceptual information, directing our attention and motivating behaviour - for example, a knock on the door, the ring of a telephone, or the screech of car tyres. Conversely, many sounds we learn to ignore though habituation, such as those caused by distant traffic or a ticking clock. Likewise, for robots to move from factory floors to mainstream environments such as domestic households or the office, they will need to be capable of classifying, recognising and discriminating between the variety of novel sounds that will inevitably occur in such environments. For robots, many sounds will be worthy of the robot's attention and should trigger an appropriate response, such as a spoken command or a fire alarm. However, as is the case with people, many sounds should be ignored by the robot.

Understandably, most research in robot perception has focused on vision systems and mapping the robot's environment, thus allowing robots to detect people, obstacles, and other important objects. Comparatively little attention been paid to developing robot auditory perceptual systems, with the notable exception of speech recognition systems. Auditory perception can allow robots to perceive important aspects of their environment that are undetectable to visual perception systems due to visual occlusion or low light, such as a knock on the door or the ring of a telephone. For robots to be useful assistants and companions in our everyday lives, the ability to hear as we do will be critical.

In this paper we present an approach for detecting, classifying and recognising novel non-verbal sounds on an Aldebaran Nao humanoid robot. Our method allows the robot to autonomously detect novel sounds, classify these sounds, and then recognise future instances. We demonstrate and test our system via an interactive human-robot game in which a person interacting with the robot can teach the robot via speech the names of novel sounds, and then test the robot's auditory classification and recognition capabilities by providing further examples of both novel sounds and sounds heard previously by the robot.

This paper is structured as follows: Section 2 outlines the benefits of developing auditory perception systems for autonomous robots. Section 3 describes common approaches to sound detection and sound classification. Section 4 presents literature related to robot hearing. In Section 5 we describe our approach, how it differs from existing approaches, and detail our implementation. Section 6 describes our experimental setup, and Section 7 describes our empirical results. Finally, in Section 8 we discuss the implications of our work, its limitations, and possible future avenues of research and development.

2 Background

Auditory perception, or "machine hearing" [1], is an emerging field of research focusing on the perception of non-verbal sounds. Machines hearing aims to endow machines with abilities such as distinguishing speech from music and background noises, to determine the direction from which a sound originates, and to learn and organise knowledge regarding sounds that have been experienced by the machine.

Sound recognition can offer a diverse range of applications and benefits for autonomous robots, from security and surveillance to health care and domestic assistance. For example, a security robot could respond to the sound of breaking glass or footsteps; a health care robot could hear a person crying or falling over and provide care; a domestic robot could respond to a knock at the door or the ring of a telephone. Coupled with auditory localisation, it is possible for robots to have active listening and thus move towards the source of a sound allowing multimodal perception.

Sound recognition can also provide feedback to the robot during physical manipulation tasks. Many common home appliances provide sound feedback to increase their ease-of-use, i.e. a successful button press results in a "click". Beeps, bells, buzzers and simple melodies are all purposefully designed as sources of feedback on many household appliances such as microwaves, dishwashers, and washing machines. Conversely, the lack of sound can also indicate a problem that requires further investigation. When a person turns they key in a car's ignition they expect to hear engine noise - the lack of engine noise is indicative of a problem with the car's engine. Surprisingly, to date there has been little leverage by robots of auditory feedback to improve task performance.

While perceiving the source of sounds comes naturally to people, for robots sound classification and recognition is not trivial. On one hand, two different objects can produce similar sounds. Conversely, real-world events from the same object can produce different sounds. Moreover, for naturally occurring sounds, the auditory signature of multiple instances of the "same" sound will have a degree of natural variation. For example, each bark from the same dog may have a similar acoustic signature, but no two barks will be identical. Furthermore, the location of the sound source relative to the robot will affect the sound's acoustic signature. Consider a stapler - the auditory signature of the stapler depends upon which part of the stapler body is depressed and on what surface is it is resting upon [3]. While human beings can easily categorise these variations in sound as belonging to the same object, for robots this a very difficult task.

3 Problem Domain

Sound recognition for robot hearing is relatively new in comparison with automatic speech recognition (ASR) and sound source localisation. In contrast to many ASR systems in which a continuous audio stream is processed, a typical acoustic event recognition system involves an audio detection module to isolate relevant sounds from background noise, and then a classification module identifies sounds from the discrete isolated sound events. This is especially the case for robotic systems, since the direct continuous analysis of the audio stream is prohibitive in terms of computational load [5]. Thus a typical acoustic event recognition system is composed of three main processes:

- Detection, which involves finding the start and end points of a sound from a continuous audio stream.
- Feature extraction, which involves extracting relevant and identifying features from an audio signal.
- Pattern classification, which compares the features of a current sound to previously trained models for identification.

3.1 Sound Detection

Detecting a sound is the cornerstone for any sound recognition and classification problem. Sound detection refers to a process which accepts a stream of audio data, generally in the form of buffers of samples and then determining from these buffers whether or not a sound is present. A sound detection process should be able to find both the start and end of a sound from these buffers. When a sound is detected in an audio buffer, the buffer will then be passed to a separate process for classification and/or recognition. Standard sound detection techniques are based on thresholding off the mean signal energy [5; 7].

3.2 Feature Extraction

One of the most common approaches from speech recognition for extracting features from an audio signal are Mel-Frequency Cepstrum Coefficients [9] (MFCCs), which represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs are also common in speaker recognition, which is the task of recognising people from their voices. Another common approach is Perceptual Linear Prediction (PLP) [10] which likewise modifies the short-term spectrum of the speech by several psychophysically based transformations which preferentially weight and threshold the original spectrum based on a model of the human audio perceptual system. These extracted features are then used to train Hidden-Markov Models [13], Gaussian Mixture Models [14], or Support Vector Machines [15]. The

learned models can then be used for detecting the learned signal in new audio data.

4 Related Work

For robots, machine hearing has the additional constraints of using the robot's on-board sensors and being processed in real-time using the robot's limited processing power. Most work related to robot hearing has focused on speech recognition and audio localisation [11]. However, there have been a few examples of audio event detection with robots. For example, Kraft et al. [8] developed a kitchen sound recognition system for a mobile robot using a novel feature extraction method based on Independent Component Analysis [16] (ICA). Their method learns ICA basis functions over a multi-frame window of features; these functions capture inter-frame temporal dependencies in an efficient manner. They compared their approach to MFCCs and found temporal ICA is a better feature set for describing kitchen sounds. However, their sound recognition system was not actually implemented on-board the robot.

Wu *et al.* [4] develop a home surveillance robot that utilises both audio and visual information. Their supervised learning approach involves the use a dataset of both "normal" sounds (e.g. speech, doors opening and closing) and "abnormal" sounds (e.g. gun shots, glass breaking, crying and groaning). Mel frequency cepstral coefficients (MFCC) are used to extract features from the audio data. The features are in turn used as input to a support vector machine classifier for analysis. They report an 83% recognition rate for abnormal sounds that are played through a speaker (as opposed to naturally occurring in the robot's environment).

Romano *et al.* [3] have released "ROAR" - ROS Opensource Audio Recognizer. ROAR is a toolkit for ROS (Robot Operating System), which allows for offline manual supervised learning of audio events. The ROAR toolkit is composed of two parts, one for learning audio events and one for detecting them. Features are extracted using Perceptual Linear Prediction (PLP) [10]. The result of the PLP processing is a set of autoregressive coefficients that encode perceptually salient information about the analyzed signal. A custom piece of software - the "ROAR Trainer" allows the user to select regions of the audio signal they wish to learn using a computer mouse. Examples are then classified using one-class support vector machines (OCSVMs). Romano *et al.* evaluate their toolkit using a stapler, drill, phone alarm and velcro. Romana *et al.* show through implementation on a robotic arm how combining contextual information with a set of learned audio events yields significant improvements in robotic task-completion rates.

Swerdlow *et al.* [6] presents a system that is trained to detect kitchen sounds using a robot with a six microphone array based on Gaussian Mixture Models in correspondence with the MFCCs as acoustic signal features. Furthermore, a Universal Background Model (UBM) is used for the special case of speaker identification. Speech data was recorded from ten speakers and five kitchen appliances (a coffee grinder, a toaster, a bread cutter, a hand-held blender, and a household electric coffee machine). They examined how the length of the training phase and the minimum data length affected recognition rates.

Janvier *et al.* [5] also enabled a Aldebaran Nao humanoid robot to detect kitchen sounds. Unlike other approaches which use MFCCs and PLPs, Janvier *et al.* use the stabilized auditory image (SAI) representation [12], which is a time-frequency sound representation close to a correlogram. The auditory images are then mapped into a vector space of reduced dimension, and a vector quantization technique is used to efficiently map the SAI representation in a feature space. An offline training phase learns prototype vectors from a set of audio data recorded with a humanoid robot in a house. The prerecorded audio consisted of 12 sound classes, each with 7 examples from 3 different positions, thus resulting in a training set of 21 examples for each of the 12 sound classes. During testing the Nao robot performed with a sound recognition accuracy rate of 91% in the worst case.

5 Our Approach

In contrast to other approaches in which the training phases occurs offline [8; 4; 3; 5] we aim to develop a system in which the training of the robot to recognise acoustic events occurs in a natural manner. Thus, necessary constraints of our approach are that the training process must occur on-board the robot, in real-time, and without the use of specialist software or external devices to facilitate the teaching of the robot. Our objective is to develop a proof-of-concept system which demonstrates that robots embedded in the home or office could learn to recognise naturally occurring environmental sounds through experience, while seeking feedback from people in the robot's environment to reinforce the learning process.

5.1 Contribution

The main contributions of our approach can be summarised as follows:

- Our system is implemented on a real robot using embedded sensors and runs on-board the robot.
- Our system's training phase occurs online, rather than offline as in the case of related works such as [8; 4; 3; 5].

Figure 1: The Nao humanoid robot. The Nao is approximately 58cm high, with 25 degrees of freedom. Picture source: http://www.aldebaran-robotics.com/

- Our system uses a small number of real examples, with the learning process involving a natural speech-based interaction process with the robot.

5.2 Hardware

We use an Aldebaran Nao humanoid robot[1]. The Nao is approximately 58cm high, with 25 DOF (see Figure 1). The Nao robot has four microphones with a frequency bandpass of 300 Hz to 18kHz. The microphones are located on the robot's head, as shown in Figure 2. The Nao robot utilises an Intel Atom Z530 processor, with a clock-rate of 1.6gHz and has 1GB RAM. The Atom processor is a low-power CPU generally used in portable devices such as netbooks and smart phones. The processing power on the Nao is directly equivalent to a Nokia Booklet 3G [18], or Dell Inspiron Mini 1210 [19], both developed in 2009 and which also make use of the Atom Z530 processor with 1GB RAM.

Sampling

The robot's microphones are sampled at 48kHz and send a buffer every 170 ms containing 4 channels (one channel for each microphone). We used the left and right microphone channels, with the front and rear microphones being ignored. The rear microphone channel was ignored due to high levels of background noise from the robot's fan, while the front microphone was ignored to reduce computational load.

[1]http://www.aldebaran-robotics.com/

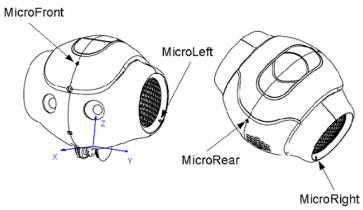

Figure 2: The Nao's four microphones. Picture source: http://www.aldebaran-robotics.com/

5.3 Sound Detection

Rather than identifying the start point and end point of a sound within the continuous buffer signal, our sound detection system simply aimed to identify buffer samples that contained sounds. This decision was made to reduce the computational complexity of searching within each buffer for the start and end point of a sound. However, a negative impact of this approach is that each buffer can contain irrelevant background noise before a sound begins and after a sound ends. To reduce the size of this negative impact, each microphone buffer was halved into two windows of 4096 samples each of 85ms duration.

To detect if there is a sound present we estimate the signal power of each window by finding the root mean square (RMS) of each window. An average signal power is then calculated for the 30 most recent windows. If the RMS value for the current window exceeds a threshold[2] compared to the average we determine that a sound is present within the current window. If sounds are detected in consecutive 85ms windows they are assumed to be part of the same sound, and thus the sound detection system identifies sequences of buffers that contain a sound.

Microphone Switching

The sound detection system switches microphones by determining whether the left or the right microphone has a higher power value. The microphone with the highest value is used by the classifier for matching. By choosing the microphone closest to the sound source we aim to process the best quality data of the two microphones. After the initial microphone determination, the same microphone is used for the remainder of the detection. For example, if a sound has a higher initial left RMS value, then for the duration of that sound the left microphone would be used.

5.4 Feature Extraction

Once a sound has been detected, the sequence of windows containing that sound are fed to the feature extractor. As described in Section 5.3, each window is

[2]Our threshold was set to be double (2 times) the average of the last 30 windows

of 85ms duration and contains 4096 samples. The feature extractor further splits each window into 8 frames of size 1024 samples with each frame overlapping 50% of the previous frame. The next step in the feature extraction process is to window each frame using Von Hann windowing [17].

The Von Hann window formula is defined as:

$$w(i) = 0.5 - 0.5 * cos(2 * \pi * i/(n-1)) \qquad (1)$$

Once the frame has been windowed, we then perform a Short-time Fourier Transform (STFT) for each frame. The STFT results in the 1024 samples per frame being reduced to 513 complex numbers. Once the STFT for each frame has been calculated, we then calculate the top 13 Mel Frequency Cepstrum Coefficients (MFCCs) for each frame (using a freely available code library for doing so [20]), further reducing the 513 frequency domain numbers into 13 numbers. Each set of 13 MFCCs for each frame is then stored in memory as a description of the frame. When all the frames for a sound have been processed the set of MFCCs for each frame are concatenated to form a feature vector describing the sound.

5.5 Training and Classification

When a new sound is detected, it is represented as a set of MFCCs, as described in Section 5.4. The classification system also treats the length (duration) of each sound as a feature, thus mapping the number (count) of the MFCCs and the actual values of the MFCCs to a feature vector. We employ a simple "class free" classification system (similar to that in [5]) in which each new sound's feature vector is stored in a list, and nothing is forgotten. Whenever a sound is detected, the classifier compares it to the known feature vectors, and searches for a match based upon Euclidean distance threshold. If a match can not be found, the robot asks via speech its human supervisor for a new category label to describe the sound (this interaction process is described in further detail in Section 6).

Threshold Tuning

To establish appropriate thresholds for determining category membership of sounds, we recorded 10 examples each of 9 different sounds from 9 different positions relative to the robot, as shown in Figure 3. For each category of sound the variance and means of MFCCs was calculated, allowing us to find suitable Euclidean thresholds separating each category of sound. It its important to note that these thresholds have then proved suitable for learning *new* sounds that were not part of this process.

6 Experimental Setup

To test the sound event classification system a simple human-robot interactive "game" was developed. This

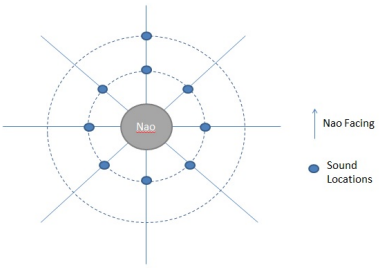

Figure 3: The sound locations used for threshold tuning.

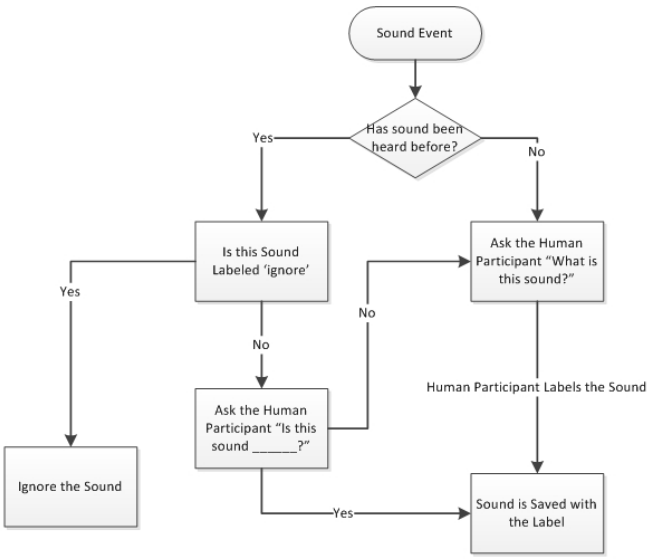

Figure 4: Flowchart illustrating the human-robot speech-based interaction sound training and testing game.

game involves the robot constantly listening for sounds. If an unknown sound is detected, the robot will ask "What is this sound?", to which the human participant can respond with the name of the sound. Conversely, if a known sound is detected the robot will ask for confirmation from the person they are interacting with that the classification of the sound is indeed correct. The human participant can then either agree with the robot, or disagree. Upon disagreement the robot will then let the human participant instruct what the sound actually is. Figure 4 describes the logic of the game process.

7 Results

Using the human-robot interaction sound training game (described in Section 6), the accuracy of the developed sound classification system was evaluated in two phases: firstly, the system was evaluated using prerecorded sounds played through a speaker; secondly, the system was evaluated using environmental sounds natu-

	Recognised Sound										
	MISS	burp	click	drill	gun	scream	smack	disconnect	ring	button3	button8
burp	11.1 %	88.9 %									
click	0.0%		100 %								
drill	11.1 %			88.9 %							
gun	11.1 %				88.9 %						
scream	0.0%					100 %					
smack	11.1 %						88.9 %				
disconnect	11.1 %							88.9 %			
ring	33.3 %								66.7 %		
button3	11.1 %									88.9 %	
button8	22.2 %										77.8 %

Table 1: Confusion matrix for 10 different prerecorded sounds played through a speaker. Training consisted of a single instance of each of the 10 sounds. After training was completed, each sound was then played 9 times. The "MISS" column represents instances in which the robot was unable to recognise the sound. Note, there was not a single example of a sound being mistaken for another sound.

	Recognised Sound					
	MISS	drawer opening	deodorant spray	mouse click	ball bounce	door knock
drawer opening	15.0 %	85.0 %				
deodorant spray	10.0 %	5.0 %	85.0 %			
mouse click	20.0 %	5.0 %		75.0 %		
ball bounce	10.0 %			15.0 %	75.0 %	
door knock	15.0 %			10.0 %	5.0 %	70.0 %

Table 2: Confusion matrix for 5 sounds naturally generated by the experimenter. Training consisted of a single instance of each of the 5 sounds. After training was completed, each sound was then played 20 times. The "MISS" column represents instances in which the robot was unable to recognise the sound.

rally generated by physical force in our laboratory (e.g. a person knocking on the door).

Table 1 displays the robot's performance in recognising 10 different prerecorded sounds. The sounds ranged from "natural" sounds (such as a person burping and screaming), to machinery noises (a gun shot, a power drill), to electronic household sounds (a phone ringing, a phone disconnect signal, and button presses). As the sounds were prerecorded, there is little variation in the acoustic signature of each sound, and as such there was not a single instance of a sound being misclassified as another sound. Accuracy was evaluated after one initial training example. The combined accuracy of the developed sound classification system is 87.78%.

Table 2 displays the robot's performance in recognising 5 different sounds generated naturally by an experimenter in our laboratory. These sounds included the bounce of tennis ball on the floor of the lab, a drawer opening, the experimenter clicking a computer mouse, an aerosol deodorant spray being pressed, and a single knock on the door to the lab. For these naturally generated sounds, due to natural variation that can occur in the acoustic signature (for example, knocking on a slightly different part of the door with a slightly different level of force with a slightly different part of the hand can generate a very different sound), the accuracy of the system dropped to 78.0%. Some similar sounds, such as the door knock, ball bounce, and mouse click, were on occasion confused by the robot.

For both prerecorded and naturally generated sounds, the majority of detection errors happen when the robot has only been exposed to two or three examples, with performance improving over time as more examples are stored in memory.

Lastly, an important aspect of the training game is a means for teaching the robot to ignore certain sounds. By instructing the robot to ignore a sound it will not respond when it hears that sound again. This was needed as when the robot turns its head to look at the participant the robot's own motors would trigger the sound detection process.

7.1 CPU and Memory Usage

Our solution was constrained by the limited processing power of Nao robot (described in Section 5.2). Peak CPU usage was measured to occur during the Short-time Fourier Transform (STFT) which results in the

1024 samples of each frame being reduced to 513 complex numbers (see Section 5.4 for implementation details). For sounds less than 1 second in duration, peak CPU usage was measured at 46.7%. When the sytem is listening for sounds (as opposed to processing and classifying them), the CPU usage was measured at 9.6%. Memory usage was relatively constant during the system's operation, hovering between 7.9% and 8.0%.

8 Discussion

We have developed an acoustic event classification system for a humanoid robot that:

- Operates in real-time.
- Requires minimal training, producing recognition rates of greater than 75% from a single training example, and with this recognition rate improving with further training examples.
- Is user friendly - the training process is driven only by human-robot speech interactions. No external software, tools or special expertise are required; anyone, regardless of their experience with robots, can teach the robot to recognise any sound that they desire.
- Allows the robot to learn to ignore irrelevant sounds as determined by the user.

Our proof-of-concept system demonstrates it is possible to develop autonomous robots capable of learning, remembering and recognising non-verbal sounds in their environment from a very small number of examples. In the future, "robot hearing" could have enormous potential in a wide variety of application domains, such as security and surveillance, household assistance, and health care. Robot hearing could also be used to generate feedback and reinforcement for the robot during visually guided tasks.

8.1 Limitations

Limitations of this work include:

- The system is only able to classify sounds which have a duration of less than one second. When a sound is detected with a longer duration, the robot will start to lose audio buffers containing the sound due to the increased computational load from continued running of both the sound detection and sound classification processes.
- Some common household sounds such as a telephone ringing or a stapler being used are composed of a sequence of two or more sounds. The detection system recognises sounds such as these as multiple individual sounds, rather than as a single collective sound.
- Lastly, the system lacks a scalable method for categorising and comparing sounds. We employ a simple "class free" approach in which each new sound's feature vector is stored in a list, and nothing is ever forgotten. Thus, if an example of a known sound is heard but the classifier fails to recognise it, the current example of the known sound is stored as a new sound in a new "category". Thus a more flexible categorisation system is required which can compare similarity between sounds, and adaptively adjust category membership boundaries.

8.2 Future Work

There are a number of immediate improvements that could be easily implemented. Most notably, implementation of a classification system that allows the robot to learn sound categories autonomously without direct human supervision. While our current process is simple and speech-based, it still requires a "human in the loop".

Furthermore, the robot requires the ability to habituate to regular but irrelevant sounds. Currently the robot needs to be told by a human supervisor which sounds to ignore, which is obviously not scalable. A better approach would be to integrate knowledge from other perceptual modalities to allow the robot to autonomously determine the relevance of particular sounds to its task.

Lastly, we intend to integrate our acoustic sound recognition system with the THAMBS attention system [2], which provides a simple but effective method for producing natural reactive behaviours in response to multimodal perceptual input.

References

[1] R. Lyon. Machine hearing: An emerging field. Signal Processing Magazine, IEEE, vol. 27, no. 5, pp. 131-139, Sept 2010.

[2] C. Kroos, D. Herath and X. Stelarc. From Robot Arm to Intentional Agent: The Articulated Head. Book chapter in Robot Arms, Satoru Goto (Ed.) InTech, DOI: 10.5772/16383.

[3] J. M. Romano, J. P. Brindza, and K. J. Kuchenbecker. ROS open-source audio recognizer: ROAR environmental sound detection tools for robot programming. Auton Robot (2013) 34:207-215.

[4] X. Wu, H. Gong, P. Chen, Z. Zhong, and Y. Xu. Surveillance Robot Utilizing Video and Audio Information. J Intell Robot Syst (2009) 55:403-421.

[5] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud. Sound-Event Recognition with a Companion Humanoid. IEEE International Conference on Humanoid Robotics (Humanoids) (2012).

[6] A. Swerdlow, T. Machmer, B. Kuhn, and K. Kroschel. Robust sound source identification for a

humanoid robot. ESSV September 2008, Frankfurt, Germany.

[7] A. Dufaux. Detection and recognition of impulsive sounds signals. PhD thesis, Facult des sciences de l'Universit de Neuchatel, 2001.

[8] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel. Temporal ICA for Classification of Acoustic Events in a Kitchen Environment. Proc. of Interspeech 2005 - 9th European Conference on Speech Communication and Technology.

[9] S. Davis, and P. Mermelstein. (1980). Comparison of parametric representations formonosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), 357-366.

[10] Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, 87(4), 1738-1752.

[11] J-M. Valin. Auditory System for a Mobile Robot. PhD Thesis. University of Sherbrooke.

[12] T. C. Walter, Auditory-based processing of communication sounds, Ph.D. dissertation, University of Cambridge, 2011.

[13] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of the IEEE, VOL. 77, NO. 2, Feb 1989.

[14] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, Jan. 1995.

[15] A. Ganapathiraju, J. E. Hamaker, and J. Picone. Applications of Support Vector Machines to Speech Recognition. IEEE Transactions on Signal Processing, VOL. 52, NO. 8, Aug 2004

[16] J-H. Lee, H-Y. Jung, T-W. Lee, S-Y. Lee. Speech feature extraction using independent component analysis. Proc. of IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2000.

[17] J. W. Picone. 1993. Signal modeling techniques in speech recognition. Proceedings of the IEEE, Vol. 81, No: 9.

[18] D. Ackerman, Review of Nokia Booklet 3G. http://asia.cnet.com/product/nokia-booklet-3g-intel-atom-z530-1-6ghz-processor-1gb-ram-44981485.htm, 2009.

[19] A. Kidman, "Dell Inspiron Mini 1210 Review". http://www.cnet.com.au/dell-inspiron-mini-1210-339295637.htm, 2009

[20] "libmfcc". C library for computing Mel Frequency Cepstral Coefficients (MFCC). Available at: https://code.google.com/p/libmfcc/source/checkout