Karlsruhe Institute of Technology

Detection of unknown sound sources for learning new object classes

Bachelor Thesis of

Jakue López Armendáriz

At the faculty of
Electrical Engineering
and Information Technology

Reviewer:	Prof. Dr.-Ing. Kristian Kroschel
Advisor:	Dipl.-Inf. Benjamin Kühn
Second advisor:	Dr. D. Luis Serrano Arriezu

Time: 21. January 2011 – 21. April 2011

Erklärung

Ich erkläre mich damit einverstanden, dass meine Studienarbeit mit dem Titel **Detection of unknown sound sources for learning new object classes** in eine Bibliothek eingestellt oder kopiert wird.

Karlsruhe, den 21.04.2011

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Studienarbeit selbständig und ohne unzulässige fremde Hilfe angefertigt habe. Die verwendeten Literaturquellen sind im Literaturverzeichnis vollständig zitiert.

Karlsruhe, den 21.04.2011

Acknowledgments

I would like to thank my parents for the moral and economical support they provided through my entire life. *Mila esker Aita eta Ama.*

I would like to show my gratitude to Prof. Dr.-Ing. Kristian Kroschel for making possible to do my Bachelor Thesis in his research group and for been the main revisor of the Thesis.

I owe my deepest gratitude to my advisor, Dipl.-Inf. Benjamin Kühn, whose expertise, understanding, and patience, make possible the writing of this thesis. I really appreciate his knowledge and his assistance during this time.

I would also like to thank my girlfriend, because she did not only give unlimited help and support, but she also brings optimism in my life.

In conclusion, I recognize that this research would not have been possible without the assistance of the Karlsruhe Institute of Technology (KIT), the Faculty of Electrical Engineering and Information Technology, the OPASCA research group, the UpNa (Public University of Navarra) and all my office colleagues.

Thanks.

Contents

1. Resumen en castellano - UpNa

Cada uno de nuestros cinco sentidos nos ayuda a interactuar con nuestro entorno. Somos capaces de oler, ver, saborear y tocar, y toda esta información que recibimos es procesada por nuestro cerebro para poder ser capaces de movernos, relacionarnos con otras personas o reaccionar ante diferentes peligros.

En el campo de los robots humanoides, todas estas características tienen que ser creadas artificialmente. Implementando diferentes algoritmos, los robots son capaces de aprender cada vez más habilidades para relacionarse con su entorno. En lo referente a percepción de imagen y sonido, el conocido como análisis opto-acústico de escenas tiene el objetivo de detectar e identificar cada uno de los eventos ópticos y acústicos que suceden en un entorno específico, con el fin de lograr una correcta percepción de este.

Con este objetivo, el Collaborative Research Center 588 - Humanoid Robots [1] fue establecido por la Deutsche Forschungsgemeinschaft (DFG) [2] y está situado en el Karlsruhe Institute of Technology (KIT) [3] desde 2001. El subproyecto P2 es el responsable de la percepción multimodal del entorno del robot, concretamente de la exploración interactiva. Por ello, un sistema de análisis opto-acústico de escenas (OPASCA - Opto Acoustic Scene Analysis) se empezó a desarrollar.

Uno de los pasos más importantes en este proceso es que el robot sea capaz de reconocer objetos y personas. Además, la información ofrecida por estos objetos y personas tiene que ser aprendida.

La localización de objetos y personas se consigue mediante el uso de un grupo de sensores multimodales. Un array de micrófonos y dos cámaras estéreo facilitan la información necesaria para realizar el proceso.

Uno de los objetivos del sistema OPASCA es establecer un lenguaje común capaz de establecer una comunicación entre robot y humano, identificando referencias ha objetos y centrando su atención en ellos, y siendo capaz de aprender la relación existente entre el aspecto visual de los objetos y la descripción facilitada por el humano.

Una percepción jerárquica y multimodal se emplea para las tareas anteriormente descritas. La percepción del entorno se debe realizar de una manera eficiente. Cuanta más información es adquirida en el tiempo, una información más detallada puede ofrecerse. Es decir, en nivel de abstracción es reducido durante la exploración. El sistema OPASCA es considerado multimodal dada su habilidad de unificar diferentes propiedades dados por diferentes modos. Los modelos multimodales son generados automáticamente, de esta manera los

objetos pueden ser reconocidos en el futuro. Dado que los entornos no son estáticos, el sistema trata de mantenerse flexible a estos cambios mediante la capacidad de añadir nueva información de cualquier objeto o persona en cualquier instante de tiempo.

Una propiedad importante de este sistema es la capacidad de aprendizaje. Para el caso concreto de nuevas fuentes de sonido, el proceso de clasificación acústica debe ser capaz de proporcionar una decisión sobre si lo que es percibido es conocido, o por el contrario, desconocido. Actualmente, el sistema OPASCA no es capaz de tomar esta decisión.

Objetivo Principal

El objetivo principal de este proyecto es desarrollar un algoritmo que posibilite el rechazo de fuentes de sonido desconocidas mediante el uso del ya existente sistema OPASCA. Para ello, será implementada una *One-Class-Support Vector Machine*. Esta tratará de ser capaz de dar una decisión sobre si lo que es percibido por los micrófonos se corresponde a alguna de las clases existentes o es desconocido.

Para conseguir un método de clasificación y rechazo de la manera más eficiente, diferentes objetos, entornos acústicos, parámetros de la señal de audio (i.e SNR) y diferentes combinaciones de representaciones paramétricas (i.e MFCC, LPC) serán utilizados.

El sistema tratará de ser lo más parecido posible a la realidad. La posición del robot humanoide no es estática, es decir, puede moverse por su entorno o incluso cambiar de escenario. Por esta razón, la decisión sobre si lo percibido es conocido o desconocido, no debe verse afectada por los cambios en el entorno. Para ello, el sistema será entrenado y probado en diferentes situaciones y emplazamientos.

El proceso puede ser separado en cuatro partes: Captura de la señal de audio, representación paramétrica, entrenamiento del modelo y testeo de los datos.

En primer lugar, es necesario un conjunto de datos para entrenar y testar el modelo. Grabaciones de diferentes aplicaciones de uso común en una cocina doméstica (i.e. máquina de café, molinillo eléctrico, cortador de pan, teléfono) deberán ser realizadas (ver Sec.5.1)

En segundo lugar, es necesaria una representación paramétrica de estos datos (ver Sec.3.1).

En tercer lugar, los datos de audio serán empleados para estimar los parámetros de los modelos acústicos (ver Sec.5.2).

Como último paso, para verificar que los modelos creados son correctos, una etapa de prueba es necesaria (ver Sec.6).

Conclusiones y trabajo futuro

Este proyecto ha desarrollado el diseño e implementación de un algoritmo de rechazo para fuentes sonoras desconocidas. Aplicaciones de cocina capaces de ser percibidas acústicamente son etiquetadas como conocidas o desconocidas.

La *One-Class-Support Vector Machine* presenta un comportamiento adecuado para realizar el rechazo deseado. El algoritmo estará integrado en el sistema OPASCA. La combinación de las representaciones paramétricas MFCC y LPC se proponen como la mejor opción para realizar esta tarea. Más concretamente, una relación señal a ruido SNR de 4,5 dB y 16 características MFCC en combinación con 6 características LPC refleja los mejores resultados.

Los resultados experimentales presentan una precisión considerablemente alta en el caso de *Cross-Validation (CV)*. Es decir, cuando los modelos para cada objeto acústico son

entrenados y probados en el mismo entorno, el rechazo de fuentes sonoras puede realizarse correctamente.

En el caso de *mismatched conditions (MM)*, cuando los modelos son entrenados y probados en diferentes entornos, la precisión del algoritmo decrece. Aunque objetos desconocidos puedan ser rechazados, algunos modelos aparecen sobre-ajustados o demasiado restrictivos, lo que supone que objetos conocidos sean clasificados como desconocidos. La precisión obtenida es menor que en caso de *Cross-Validation (CV)*.

La captura de datos acústicos en cortos periodos de tiempo muestra unos resultados correctos. Cuando más datos son capturados en el tiempo, la desviación estándar decrece, lo que se traduce en una mayor estabilidad del algoritmo.

Para un trabajo futuro, podrían mejorarse los resultados para el caso de *mismatched conditions (MM)*. Por ejemplo, podría ser introducido un entorno o habitación modelo para evitar la dependencia a diferentes espacios.

Además, dado que en los diferentes espacios el ruido de fondo es un factor influyente, este podría ser añadido a las señales de audio para hacer los modelos más realistas.

Como punto final, para probar la robustez del algoritmo clasificador implementado, podrán ser añadidas diferentes aplicaciones comunes en una cocina. La habilidad de rechazar fuentes de sonido desconocidas abre la puerta al aprendizaje de nuevos objetos para el robot. La creación de modelos sin supervisión externa facilitará una mejor y más rápida adaptación del robot humaniode en diferentes escenarios.

2. Introduction

Every one of our 5 senses helps us every moment to interact with the environment. We are able to enjoy food because we can smell and taste, we can skip objects while walking because we can see them and it hurts when a needle stings because we can touch. All the information we receive is processed by our brain for been able to move, talk to other people or jump out when a car is approaching to us.

In the area of humanoid robots, all this features have to be artificially created. That means, all the humanlike perceptive skills have to be developed. By implementing different algorithms, more and more abilities are added and the robot starts to be able to interact with the environment. Concerning to image and sound perception, the so called opto-acoustic scene analysis has the goal of detecting and identifying every acoustic and optic event happening in a specific environment, in order to be able to achieve a good perception of it.

For this purpose, Collaborative Research Center 588 - Humanoid Robots [1] has been established by the Deutsche Forschungsgemeinschaft (DFG) [2] and is located at the Karlsruhe Institute of Technology (KIT) [3] since 2001. The subproject P2 is responsible for the multimodal perception of the environment of the robot and specially for the interactive exploration. Therefore, a opto-acoustics scene analysis system (OPASCA) has been developed.

One important step is that, the robot should be able to recognize the known persons and objects. Additionally, information about unknown persons and objects has to be learned.

Localization and classification of persons and objects is achieved with the information given by a group of multimodal sensors. A microphone array and two stereo camera will provide the necessary information to perform the process.

Some of the challenges of the OPASCA system are to set a common language to allow communication between robot and human, identify object references and focus his attention on them, and to learn the relation between the oral description given by a human and the visual appearance of an object.

A hierarchical multimodal perception structure is followed for the above explained tasks. The perception of the environment has to be performed in an efficient way. When more data over time is acquired, more detailed information can be given, that is, the level of abstraction is reduced during the exploration. The OPASCA system is also considered multimodal due to the ability of unifying different properties given by different modalities.

Multimodal models are generated automatically, so that objects can be recognized in the future. As environments are not static, that means, they are continuously changing, the system trys also to stay flexible to changes by been able to add new information of any object or person at any time.

An important property of the system is its ability to learn. The reason for this is the possibility to lead the system to new objects.

For learning new sound sources, the acoustical classification needs to be able to give a decision if the tested is known or unknown. Nowadays, the OPASCA system is not able to give this decision.

The main goal of this thesis is to develop the possibility of rejecting unknown sound sources using the already existing OPASCA system. Furthermore, a One-Class-Support Vector Machine will be implemented. We will try to give a decision about what is captured by the microphones fits one of the current model classes or, in the other hand, it is unknown. In order to find the most efficient method for rejection and classification, different objects, rooms, audio signal parameters (e.g SNR) and combinations of parametric object representations (e.g MFCC, LPC) will be used.

The system trys to be as accurate as possible to reality, and the humanoid robot's position is not static, that means, it could move between different locations or even change the scenery. For this reason, the decision between known or unknown has not to be affected by this environment changes. That is why, the system will be trained and tested in different environments.

The process can be separated in to four parts, audio recording, parametric representation, model training and data testing. First of all, a training data set is necessary for training and testing. Recordings of different appliances (e.g. coffee machine, blender, bread cutter, telephone) have to be done. This data has to be represented in a parametric way, so we will ned to find the different representations that could achieve this. In the model training segment, the training audio data is used to estimate the parameters of the acoustic models. And finally, for verifying if the models created are performing correctly, a testing stage is needed.

3. Acoustic Scene analysis

3.1. Acoustic Signal Parametric Representations

In order to train and test the classification and rejection systems, the selection of the best representation is an important task, so that the sound signal can be converted to an appropriate parametric representation.

The main goals in the selection of a parametric representation are: Firstly, compress the audio data by eliminating information not related to the acoustic characteristics of the sound source. When a significant amount of reference information is stored, such as different appliances sound signals, compact storage of the information has to be taken into consideration. Secondly, enhance those aspects of the signal that could contribute significantly to classification and rejection of the appliances [4].

3.1.1. Mel Frequency Cepstral Coefficients

With the purpose of converting sound signals to some type of parametric representation, **M**el **F**requency **C**epstral **C**oefficients are used. This features provide an alternative representation of the spectrum which incorporates the known variation of the human ear's critical bandwidths with frequency. Each step in the process of creating the features is motivated by perceptual or computational considerations. A examination of the steps is done in the next paragraphs.

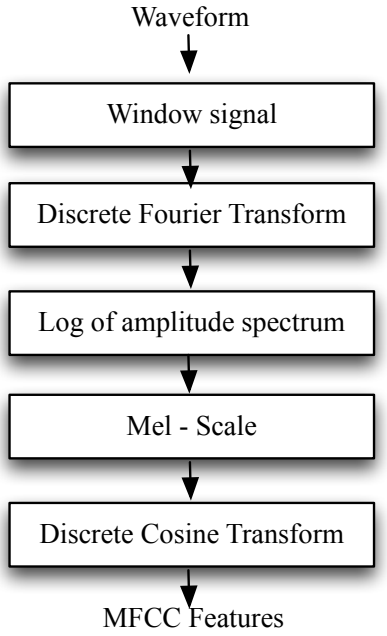


Figure 3.1.: Process to create MFCC Features

The first step is to window the input data, so that the sections of small data are statistically stationary. Normally a hamming window is used for removing edge effects.

The second step is to calculate the **D**iscrete **F**ourier **T**ransform (DFT) of each of the windows of the signal. Information about phase is discarded and only the logarithm of the amplitude spectrum is taken. The logarithm is performed because the perceived loudness of a signal has been found to be approximately logarithmic [5].

The next step is the so called Mel-Frequency wrapping. As the human auditory system does not perceive pitch in a linear way, the Mel scale is based on a mapping between each actual frequency and the perceived pitch. Each frequency is calculated by following the equation:

$$Mel(f) = 1127ln(1 + f/700) \tag{3.1}$$

The mapping is almost linear bellow 1kHz and logarithmic above this frequency. In the practical case, as we are more interested in the envelope of the frequency response instead of the frequency response itself, we use the triangular bandpass filters (see Fig.3.2) to achieve the mapping. The positions of these filters are equally spaced along the Mel-frequency.

The resulting output components of the Mel-spectral vectors appear to be highly correlated. Therefore, in order to reduce the number of parameters, a decorrelation must be done. Theoretically, the decorrelation of the vector components is achieved by **K**arhunen-**L**oeve (KL) transform [5]. For the last step, in practical cases, the KL transform is approximated by the **D**iscrete **C**osine **T**ransform (DCT). The first component of the feature vector represents the average energy of the analyzed segment, consequently it is not used in the analysis.

Finally, the sound signal of each appliance is characterized by vectors of MFCC, which will represent the acoustical features of each object.

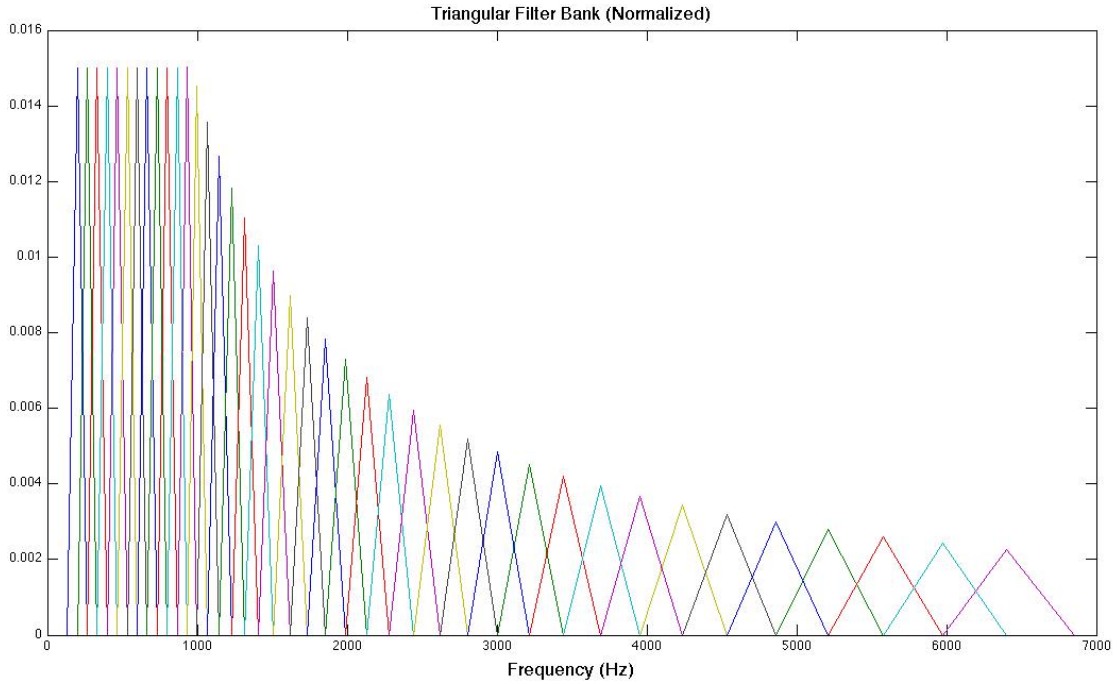


Figure 3.2.: Normalized triangular filter bank for frequency wrapping

3.1.2. Linear Prediction Coefficients

Linear Prediction Coefficients are also used for the parametric representation of the sound sources. The popularity of Linear Predictive Coding derives from its compact yet precise representation of the spectral magnitude as well as the relatively simple computation [6].

In the standard formulation of linear prediction, the model parameters are selected to minimize the mean-squared error between the model and the acoustic data. The alternative is being used for performing this linear prediction, the autocorrelation method, the minimization is carried out for a windowed segment of data. In the autocorrelation method, minimizing the mean-square error of the time domain samples is equivalent to minimizing the integrated ratio of the signal spectrum to the spectrum of the all-pole model. Figure 3.3 shows an overview of the whole approach and is described subsequently.

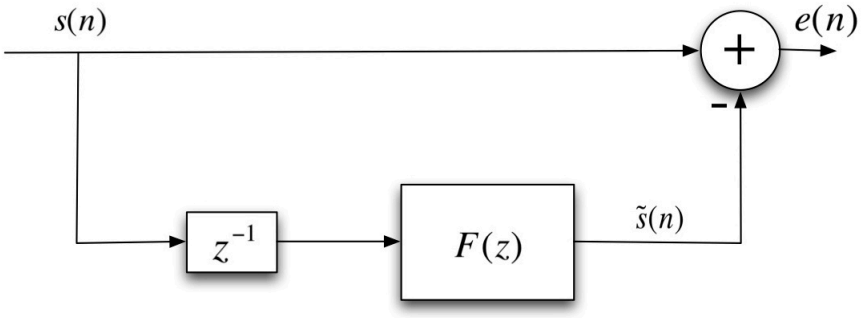


Figure 3.3.: General overview of the Linear Prediction Process

Given a signal $s(n)$, consider the problem of predicting the current value from the previous value,

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.2}$$

This prediction will produce an error by some amount:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.3}$$

The main goal will be to minimize the error by finding the optimal value of $\{\alpha_k\}$. For this, short-time average prediction error is defined:

$$\begin{aligned} E &= \sum_n e^2(n) \\ &= \sum_n \{s(n) - \sum_{k=1}^{p} \alpha_k s(n-k)\}^2 \\ &= \sum_n s^2(n) - \sum_n \{2s(n) \sum_{k=1}^{p} \alpha_k s(n-k)\} + \sum_n \{\sum_{k=1}^{p} \alpha_k s(n-k)\}^2 \\ &= \sum_n s^2(n) - 2\sum_{k=1}^{p} \alpha_k \sum_n s(n)s(n-k) + \sum_n \{\sum_{k=1}^{p} \alpha_k s(n-k)\}^2 \end{aligned} \tag{3.4}$$

The error with respect to α_l for each $1 \leq l \leq p$ can be minimized by differentiating E and setting the result equal to zero:

$$\frac{\partial E}{\partial \alpha} = 0 = -2\sum_n s(n)s(n-l) + 2\sum_n \{\sum_{k=1}^{p} \alpha_k s(n-k)\} s(n-l) \tag{3.5}$$

rearranging terms:

$$\sum_n s(n)s(n-l) = \sum_{k=1}^{p} \alpha_k (\sum_n s(n-k)s(n-l)) \tag{3.6}$$

or,

$$c(l,0) = \sum_{k=1}^{p} \alpha_k c(k,l) \tag{3.7}$$

This equation is know as the linear prediction 'Yule-Walker' equation. $\{\alpha_k\}$ are known as *Linear Predictor Coefficients*. By enumerating the equations for each value of l, we can express this matrix form:

$$\bar{c} = \underline{C}\bar{\alpha} \tag{3.8}$$

where,

$$\bar{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} \underline{C} = \begin{bmatrix} c(1,1) & c(1,2) & \dots & c(1,p) \\ c(2,1) & c(2,2) & \dots & c(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ c(p,1) & c(p,2) & \dots & c(p,p) \end{bmatrix} \bar{c} = \begin{bmatrix} c(1,0) \\ c(2,0) \\ \vdots \\ c(p,0) \end{bmatrix} \tag{3.9}$$

The solution to this equation involves a matrix inversion and it is known as the *covariance method*

$$\bar{\alpha} = \underline{C}^{-1}\bar{c} \tag{3.10}$$

Using a different interpretation of the limits on the error minimization, by forcing data only into the frame to be used, we can compute the solution it's being used in the algorithm, the linear prediction equation using the *autocorrelation method*:

$$\bar{\alpha} = \underline{R}^{-1}\bar{r} \tag{3.11}$$

where,

$$\bar{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} \underline{R} = \begin{bmatrix} r(0) & r(1) & \dots & r(p-1) \\ r(1) & r(0) & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & r(0) \end{bmatrix} \bar{r} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix} \tag{3.12}$$

$\underline{R}$ is symmetric and all the elements of the diagonal are equal so an inverse always exists. The linear prediction process can be represented as a filter by:

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.13}$$

and

$$E(z) = S(z)A(z) \tag{3.14}$$

where

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{3.15}$$

$A(z)$ is called the analyzer. The expression of the error can be represented as:

$$E = \sum_n e^2(n) = \sum_n \{s(n) - \sum_{k=1}^{p} \alpha_k s(n-k)\}^2 \tag{3.16}$$

Substituting the expression for $\{\alpha_k\}$ shows:

The *autocorrelation method*:

$$E = r(0) - \sum_{k=1}^{p} \alpha_k r(k) \tag{3.17}$$

It is also important to take into account the order of the model. For speech, if the prediction order is too small, the formant structure is not well represented. If the order is too large, pitch pulses as well as formants start to be represented. Tenth-order or twelfth-order analysis is typical for speech [6, 7].

3.1.3. Linear Prediction Cepstral Coefficients

Once LPC has been obtained, it is possible to derivate a different parametric representation of the acoustic signals from it. By applying the cepstrum to the above explained LPC features, the so called **L**inear **P**rediction **C**epstral **C**oefficients (LPCC) are obtained.

For obtaining this features, we perform the same process as for obtaining the common LPC but finally one step more is performed: The cepstrum of the LPC features is calculated.

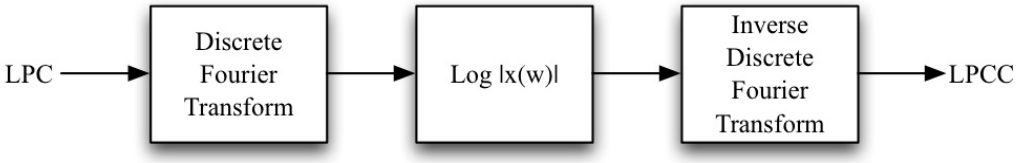


Figure 3.4.: General overview of the Linear Prediction Cepstral Process

For calculating the cepstrum, first, the **D**iscrete **F**ourier **T**ransform (DFT) of the coefficients must be done. After this, the logarithm of the magnitude spectrum is taken and, finally, the **i**nverse textbfDiscrete **F**ourier **T**ransform (iDFT) concludes the process.

Compared to LPC, the main advantage of the cepstral coefficients is that they are approximately decorrelated.It is important to note that while there are a finite number of LPC coefficients, the number of cepstrum coefficients are not [8, 9] . One of the options for taking a finite number of coefficients is to follow the order of linear prediction.

3.2. Influence of room acoustics

One of the goals of the OPASCA system is to adjust to real environments as much as possible. The challenge of the classification and rejection task when the robot operates in several rooms has to be also taken into account.

It is know that different environments could have big influence in the results. Not only direct sound is acquired by the microphones, but also a high number part of the signal which have been reflected. When a wave reaches the boundary between one medium another medium, a portion of the wave undergoes reflection and a portion of the wave undergoes transmission across the boundary. The amount of reflection is dependent upon the dissimilarity of the two media. A hard material such as concrete is as dissimilar as can be to the air through which the sound moves; subsequently, most of the sound wave is reflected by the walls and little is absorbed.

One of the acoustical properties that have more influence in the acoustics of a room is the **R**everberation **T**ime (RT). Due to the different dimensions and materials placed in the rooms (e.g walls, chairs, computers, windows, wall and floor materials...), RT will change in every of them. It represents the time in seconds that it takes for sound reflections within a space to become inaudible after the presence of a sound. It's also one of the most basic indicators of the sound quality within a space. Short RT's (< 1 sec) are preferred for high quality intelligibility, whereas long RT's (> 1.5 sec) are preferred for music listening. The simple Sabine decay formula is a classic derivation of the RT [10]:

$$RT = 0,1611 \frac{V}{S \cdot a} \tag{3.18}$$

where V is the volume of the room in m^3, S total surface area of room in m^3, a is the average absorption coefficient of room surfaces, and the product Sa is the total absorption in Sabins.

Thus, the RT is proportional to the volume of the room, and inversely proportional to the amount of absorbing material within the space. For instance, a small office with a low ceiling and carpet will have a short RT, whereas a large room like an gymnasium with hardwood floor will have a longer RT.

To find the practical impact of the room in room acoustics, investigations of room acoustic impulse responses using convolution techniques have been done [11]. The rooms acts like a big filter, selectively intensifying some sounds, softening others, and spatially scrambling the sound sources. Considering the room a linear time-invariant system with impulse response h(t):

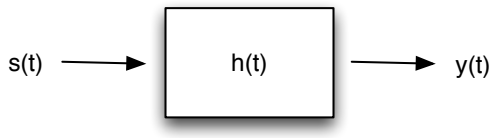


Figure 3.5.: Simplified room system

$$y(t) = s(t) * h(t) \tag{3.19}$$

or

$$y(t) = (s * h)(t) = \int_{-\infty}^{\infty} s(t) \cdot h(t - \tau) d\tau \tag{3.20}$$

where $s(t)$ is a sound that is recorded in an anechoic room (dry recording) and played back in a standard room, $h(t)$ the impulse response of the reverberant room and $y(t)$ the convolved sound as it is been recorded in that specific room.

It has to be noted that the consideration of a linear time-invariant system is just a general way of representing the environment influence problem. The estimation of the impulsive response of a room becomes more complex. Minor changes like opening a window, movements of the objects or people inside the room, the position of the source in the room and even temperature and humidity affect directly to the impulse response of each environment.

3.3. Multi Class Classification

The current implementation of the acoustic classification algorithm in OPASCA is based on **G**aussian **M**ixture **M**odels (GMM) and **M**el **F**requency **C**epstral **C**oefficients (MFCC) as acoustic features. For the case of speaker identification a **U**niversal **B**ackground **M**odel (UBM) is used.

3.3.1. Gaussian Mixture Models

The required statistical model for sound source classification in the system is created by Gaussian Mixture Models, one of the so called unsupervised classifiers, as the training samples are not labeled to show their category membership [12]. GMM trys to estimate the **p**robability **d**ensity **f**unctions (pdf) of the given observations. The conditional probability density functions of the observation vector with respect to the other classes are modeled as a linear combination of multivariate Gaussian pdf's. Each of the Gaussian follows the general form [13]:

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\sum|^2} e^{[-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)]} \tag{3.21}$$

- x is a d-component feature vector.
- μ is the d-component vector with the mean of each feature and
- $\sum$ is the d-by-d covariance matrix and $|\sum|$ its determinant. It represents the dispersion of the data on the d-dimensions of the feature vector. The diagonal elements of the matrix are the variance of x, and the non diagonal elements covariances between features. $\sum$ is diagonal if we make the assumption that the features are independent and $p(x)$ can be written as the product of the univariate probability densities for the elements of x:Each multivariate Gaussian pdf is completely defined if we know $\theta = [\mu, \sum]$

The OPASCA system only uses diagonal covariance matrices, resulting in a higher computational efficiency. Empirical investigations show that diagonal-matrix GMMs normally outperform full matrix GMMs [14].

Extracting information from a unlabeled data set can only be possible if certain assumptions are made[12]. The assumptions are the following:

- The samples come from a known number of classes
- The a priori $P(w_j)$ probabilities for each class w_j are known
- The form of the class-conditional probability densities $p(x|w_j, \theta_j)$ are known for all classes, $j = 1 \dots c$ (there are a sum of K multivariate gaussian probability functions)
- The values of the c parameter vectors $\theta_{j=1\dots c}$ are unknown (the weights of the N gaussian pdf's, the mean vector and the covariance matrix for each class)

For the training of the GMM, we consider a set X of m observations of d features: $X = [x_1, x_2 \dots x_m]$

Assuming that the observations are independent and identically distributed, the likelihood that the entire set has been produced by class C_0 is:

$$p(X|C_0) = \prod_{i=1}^{m} p(x_i|C_0) \tag{3.22}$$

Each $p(x_i|C_0)$ is modelled as a mixture of K multivariate gaussians:

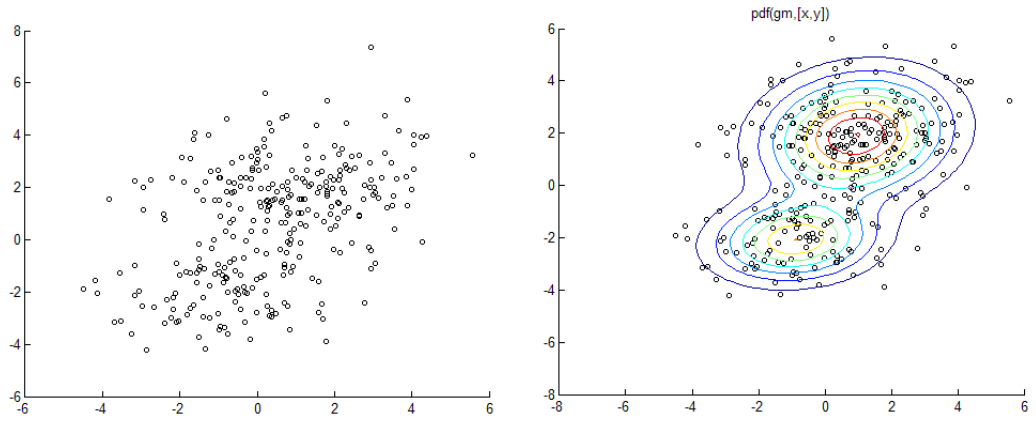
$$p(x_i|C_0) = \sum_{l=1}^{K} P(l|C_0) \cdot p(x_i|l, C_0) \tag{3.23}$$

where $p(x_i|l, C_0) = N(\mu_{l,0}, \sum_{l,0})$ is the probability of x_i being produced by the gaussian of index l in the sound source class 0. On the other hand, $P(l|C_0)$ is the prior probability of having a gaussian l for the sound source class 0. It is a weight that changes with the sound source class.

3.3.2. Expectation-Maximization

As mentioned before, for the classification of all the sound source classes a training phase is needed. In this step, the estimation of the GMM parameters $P(l|C_i)$, $\mu_{l,i}$ and $\sum_{l,i}$ with $l = 1 \dots K$ must be done.

The ideal way of approaching them would be the **M**aximum **L**ikelihood **E**stimation (MLE). Theoretically, MLE consist of finding $C = [C_{i1}, C_{i2} \dots C_{iK}]$, maximizing $P(X|C_i)$, the likelihood of observing X as being produced by the the sound source class i. That is, we



(a) (b)

Figure 3.6.: (a) Example of random training data. (b) Estimated probability density function for a two-component mixture distribution.

wish to estimate the model parameters for which the observed data are the most likely. In the case where all the parameters are unknown, MLE method becomes very complex [13].

Due to this issue, one of the very often used solutions, **E**xpectation-**M**aximization (EM) algorithm, [15] is implemented. The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. It receives this name because each iteration of the algorithm consists of an expectation step (E-step) followed by a maximization step (M-step).

In the E-step, the missing data is estimated given the observed data and current estimate of the model parameters. In the M-step, the likelihood function is maximized under the assumption that the missing data is known. Instead of the actual missing data, the estimation of the missing data from the E-step is used [16].

To explain the process, the *log likelihood function* is introduced.

$$L(C) = lnP(X|C). \tag{3.24}$$

The likelihood function is considered to be a function of the parameter C given the data X. Since $ln(x)$ is a strictly increasing function, the value of C which maximizes $P(X|C)$ also maximizes $L(C)$. The EM algorithm is an iterative procedure for maximizing $L(C)$. Assume that after the n^{th} iteration the current estimate for C is given by C_n. Since the objective is to maximize $L(C)$, we wish to compute an updated estimate C such that,

$$L(C) > L(C_n) \tag{3.25}$$

Equivalently we want to maximize the difference,

$$L(C) - L(C_n) = lnP(X|C) - lnP(X|C_n). \tag{3.26}$$

Its important to take account of the training set provided to the GMM. This has to be well thought out so that the model can be general and representative enough for all the sound source classes in the set.

3.3.3. Universal Background Model

For the special case of speaker classification, The GMM modeling is extended by a so called Universal Background Model. It has been the basis of the top performing systems since 1996 [17]. UB models are a GMM-based system developed by MIT Lincoln Laboratory. Using UBM technique, independent-speakers are modeled by only one large GMM general class. The general model is trained with speech samples from a large set of speakers to represent general speech characteristics.

Instead of applying the same method as for the appliances, each speaker individual models are derived form the UBM. A form of Bayesian adaptation in combination with speaker-specific training vectors is used. The basic idea in the adaptation approach is to derive the specific model by updating and adapting the well-trained parameters in the UBM. This provides a tighter coupling between the speakers model and UBM which, not only produces better performance than decoupled models but also allows a fast-scoring technique.

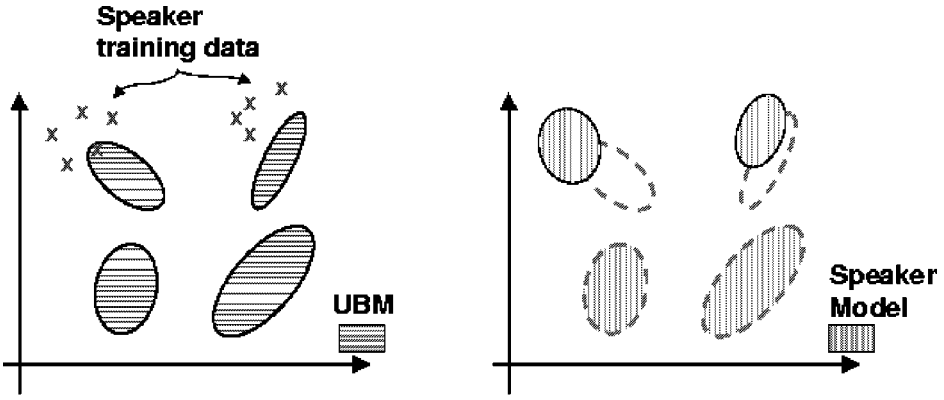


Figure 3.7.: Training vectors mapped and adapted to the UBM. Extracted from [17]

Like the EM algorithm, the adaption is a two step estimation process. The first step is identical to the expectation step of the EM algorithm, where estimates of the sufficient statistics of the speaker's training data are computed for each mixture in the UBM. The second step differs from the EM algorithm, for adaptation these new sufficient statistic estimates are then combined with the old sufficient statistics from the UBM mixture parameters using a data-dependent mixing coefficient. For final parameter estimation, the method is designed so that mixtures with high counts of data from the speaker rely more on the new sufficient statistics and mixtures with low counts rely more on the old sufficient statistics.

The training vectors ($x's$) are probabilistically mapped into the UBM mixtures. The adapted mixture parameters are derived using the statistics of the new data and the UBM mixture parameters. The adaptation is data dependent, so UBM mixture parameters are adapted by different amounts.

4. Rejection of unknown sound sources

So far, the OPASCA system is able to classify different sound sources by using the before mentioned GMM and UBM model implementations. The classification is done for a closed data set, this means, if some appliance or sound not contained in this set is perceived and processed by the system, a probability of how similar this input sound is between the known appliances will be given. Namely, we will get a value in which the result will express the similarity with the modeled sounds. It can be seen that OPASCA system is not able to give a decision about what it's been perceived it's known or unknown.

Henceforth, the main goal will be to find a method for the rejection of unknown sound sources, in other words, to find a method for identifying appliances out of the closed data set.

4.1. Adding a new outlier class

One of the first approaches for performing the rejection option will be to improve the GMM based existing multi class classifier by adding a new rejection or unknown class. Providing to the training phase of the model an outlier class training data set will result in a uniform distribution of the unknown class in the feature space of the model. Theoretically this will allow to perform the rejection.

We find a problem when we realize that it is not possible to obtain a representative training data set. Every sound source that it is not in the actual model should be considered as an outlier. When our data set is a closed set composed by a few sound sources, and even if the close set is bigger, the number of unknown sources still shows unlimited. Building a data set, even with artificial generation, which could represent all this unseen class appears unachievable.

The incapacity of building a good training data set will not allow to train properly the GMM model. Thus, the missing of outlier training data forces us to find a different solution for rejection.

4.2. Standard approach to rejection

The standard approach to distinguish the outlier class from the rest of the known classes, is to set a threshold t_d in the total data density. The rejection is based on thresholding the posterior probabilities obtained in the test. This approach is known as 'Chow's rule'

[18]. Because each model characterizes the same outlier class with their threshold t_{di}, these thresholds should coincide. Been w_i the i number of w classes, according to Chow's rule a pattern x is rejected if:

$$\max_{k=1,\ldots,N} P(w_k|x) = P(w_i|x) < t_d, \tag{4.1}$$

where $t_d \in [0, 1]$. On the other hand, the pattern x is accepted and assigned to the known class w_i, if

$$\max_{k=1,\ldots,N} P(w_k|x) = P(w_i|x) \geq t_d, \tag{4.2}$$

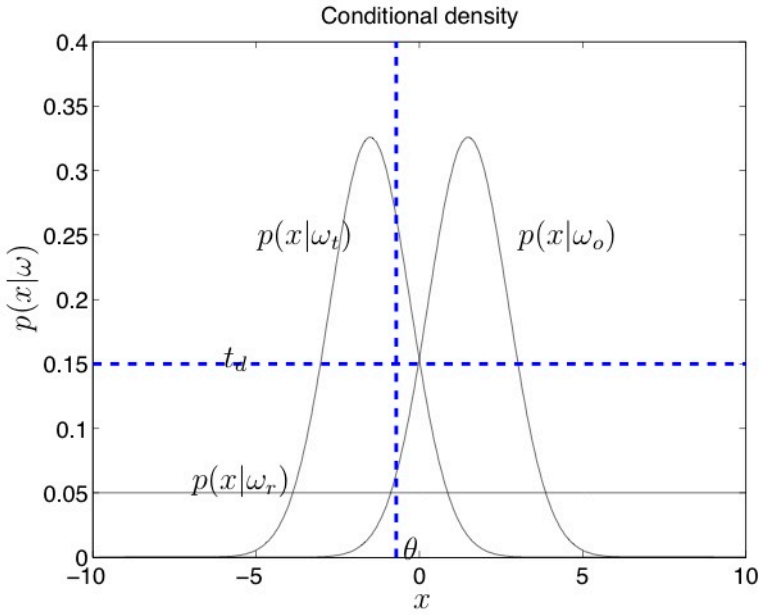


Figure 4.1.: Illustration of the class conditional densities for w_t, w_o and w_r. Classification boundary specified by θ and rejection boundary by t_d. Extracted from [19]

Because the objects in the feature space appear overlapped, the outlier objects do not typically appear in areas with a low posterior probability (i.e. areas between the known classes), but they are often distributed around the known classes. Here the total data probability density is low, but the posterior probabilities are high [20].

Due to this issue, outlier objects will be considered part of the closed data set model. A simple threshold shows to be not robust enough for performing the rejection step.

4.3. Combining Classifiers

For rejecting examples occurring far away from the sample class, the limitation of the reject option approach is that a model chosen for good classification performance does not necessarily imply good rejection performance, and vice-versa. If the same model is used for classification and rejection, we may have to give preference to the performance of one of the classifiers [21].

One classification strategy that could avoid this problem consist of a sequential combination of a one-class and a multi-class classifier. The proposed two stage process allows both rejection and classification performance to be adapted specifically for improving the

respective models and representations. That is, a classifier can be designed to obtain good performance on unknown classes, and a separate classifier model can be set to perform the multiple class classification.

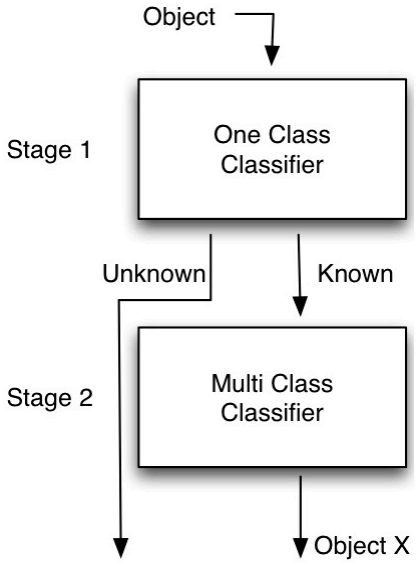


Figure 4.2.: General view of the tow stage classification process

The input sound source will be first tested in the one class classifier to decide if it fits to some of the models of the closed data set. If the output of the classifier is a positive case, in other words, when the source is labeled as known, the next step consists of performing the multi class classification, so that we can obtain the information about which sound source of the data set is perceived.

As mentioned before, the OPASCA system is already implementing a GMM based multi class classifier, so the main goal will be to develop a one class classification stage.

4.4. One Class Classification

The problem of classification could be explained as the way to assign a new object to one of a set of classes which are known before. The classifier which should perform this classification operation is based on a set of example objects. In the case of one class classification, an object should be classified as a genuine object (from our data set), or as an outlier object (out of the data set).

The one-class classification problem differs in one essential aspect from the conventional classification problem. In one-class classification it is assumed that only information of one of the classes, the target class, is available. This means that just example objects of the target class can be used and no information about the other class of outlier objects is present.

This means, that the boundary between the two classes has to be estimated only from data of the known class. The task is to define a boundary around the target class, such that it accepts as many of the target objects as possible, while it makes minimum the possibility of accepting outlier objects.

Three main approaches can be distinguished for the one class classification problem: The density estimation, the boundary methods and the reconstruction methods. For each of the three approaches, different concrete models can be constructed.

In all one class classification problem two distinct elements can be identified. The first element is a measure for the distance $d(z)$ or probability $p(z)$ of an object z to the target class represented by the training set X_{train}. The second element is a threshold θ on this distance. Incoming objects are accepted by the classifier when the distance to the target class is smaller than the threshold θ_d:

$$f(z) = I(d(z) < \theta_d) \tag{4.3}$$

or when the probability is higher than the θ_d:

$$f(z) = I(p(z) > \theta_d) \tag{4.4}$$

where I is the indicator function defined as:

$$I(A) = \begin{cases} 1, & \text{if A is true} \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

The one-class classification methods differ in their definition and optimization of $p(z)$ or $d(z)$ and thresholds with respect to the training set X_{train}. The most important feature of one-class classifiers is the tradeoff between the fraction of the target class that is accepted, f_{T+}, and the fraction of outliers that is rejected, f_{O-}. The f_{T+} can easily be measured using an independent test obtained from the same target class source. To measure the f_{O-} on the other hand, an outlier density, anything out of the target class, even if can also be tested by randomly choosing some objects not on the set, has to be assumed.

To compare different one-class classification methods, not only the f_{T+} and f_{O-} are important, but also other features:

As it has been assumed that the training set is a characteristic representation of the target distribution, the method should have robustness to outliers. Objects from the target set should be as much as possible accepted, and outliers should still be rejected.

One of the most important aspects for easy operation of a method by the user, is the number of free parameters that have to be chosen beforehand, as well as their initial values. When a large number of free parameters is involved, finding a good working set might be very hard. This becomes even more prominent when the parameters involved are not intuitive quantities which can be assumed, derived or estimated a priori. When they are set correctly, good performances will be achieved, but when they are set incorrectly, the method might completely fail. In some cases, these numbers cannot be intuitively given beforehand, and only by trial and error a reasonable combination can be found.

Computation and storage has also to be taken into account. Although computers are more powerful and have more storage capacity every day, methods which require several minutes for the evaluation of a feature vector might be unusable in practice. Training is not often done in real time, however, as the idea will be to develop a method for a changing environment, it could happen that it is also done in real time, that is why these training costs have also to be taken into account.

4.5. One-class SVM

Between all the choices of one-class classification methods, **S**upport **V**ector **M**achine (SVM) is chosen. This method was introduced by Vapnik [22]. A one-class Support

Vector Machine is chosen due to the capacity of directly obtaining the boundary around a target class, which is specified by a training data set.

In the most simple case of one-class SVM, a hypersphere containing all target objects is computed. To minimize the chance of accepting outliers, the volume of this hypersphere is minimized. It also offers the ability to map the data to a new, high dimensional feature space without much extra computational costs. By this mapping more flexible descriptions than the sphere are obtained. As a simple example this simple hypersphere case is explained in next paragraphs [23].

The sphere is characterized by a center a and radius R. The main goal will be that the sphere contains all the objects from the given training set $X_{train} = \{x_1, \ldots, x_N\}$. When this requirement becomes true, the empirical error is set to 0. Defining the structural error as:

$$\epsilon_{struct}(R, a) = R^2 \tag{4.6}$$

which has to be minimized by:

$$\|x_i - a\|^2 \leq R^2, \quad \forall_{i=1,\ldots,N} \tag{4.7}$$

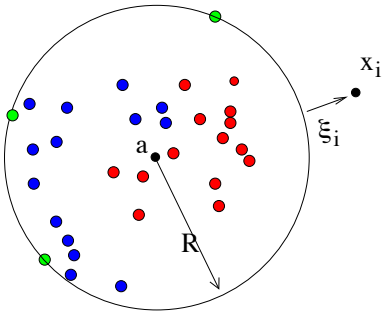


Figure 4.3.: The hypersphere containing the target data, described by the center a and a radius R. The three data points define the boundary, they are called *support vectors*. Data point x_i is considered an outlier. ξ represents a slack variable. Extracted from [23].

Slack variables $\xi, \xi_i \geq 0 \forall_i$, which represent points that lie out of the sphere, that is, empirical error, are introduced:

$$\epsilon(R, a, \xi) = R^2 + C \sum_i \xi_i \tag{4.8}$$

which constrains that almost all points are inside the sphere:

$$\|x_i - a\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall_{i=1,\ldots,N} \tag{4.9}$$

The parameter C gives the tradeoff between the volume of the description and the errors. The free parameters, a, R and ξ, have to be optimized, taking the constraints (4.9) into account. The minimization of this error with the constraints is a well-known quadratic programming problem.

We can define the center of the sphere as a linear combination of points with weights α_i:

$$a = \sum_i \alpha_1 x_i \tag{4.10}$$

Therefore, for the computation of a, points with 0 weight ($\alpha_i = 0$) can be discarded. Only points with positive weight $\alpha_i > 0$ are needed in the description of the data set. It is shown that in the minimization of (4.7), often a large fraction of the weights becomes 0. The sum in equation (4.10) is then over just a few objects x_i with non-zero α_i. These objects will be called the ***S**upport **V**ectors* (SVs) of the description.

Because we are able to give an expression for the center of the hypersphere a, we can test if a new input object z is accepted by the description. For doing this, the distance from the object z to the center of the hypersphere a has to be calculated. A test object z is considered part of the target class when this distance is smaller than or equal to the radius:

$$\|z - a\|^2 \leq R^2 \tag{4.11}$$

By definition, R^2 is the (squared) distance from the center of the sphere a to one of the *support vectors* on the boundary.

We can define now the one-class SVM classifier as:

$$f_{\text{SVM}}(z, R) = I(\|z - a\|^2 \leq R^2) \tag{4.12}$$

Where the indicator function I is defined by (4.5).

The hypersphere is a very rigid model of the boundary of the data. In general, it cannot be expected that this simple approach for the model will fit the data correctly. By mapping the data into a new representation, we could obtain a better fit between the actual data boundary and the hypersphere model. For so, a mapping of the data Φ is introduced. In this formulation, the mapping Φ is never used explicitly, but it is only defined implicitly by the kernel K.

This technique, to map the data into a new feature space, is found by Vapnik [22] and it is known as the *kernel trick*. As said, the data is mapped to another feature space where it is linearly separable. This trick has also the advantage that the introduction of the kernels does not introduce much extra computational costs. The optimization problem remains identical in the number of free parameters. The only extra cost is in the computation of the kernel functions $K(x_i, x_j)$.

We stress on the difference between the feature space, which is a space of functions, and the space of feature vectors, which is $\mathbb{R}^d$. Confusion between these two spaces is possible, we will refer these names as they are widely used in the literature.

Many different kernel functions have been proposed for SVM [24]. The most commonly used kernel function is the so called Gaussian kernel [25] and is given by:

$$K(x_i, x_j) = exp\left(-\frac{\|x_i, x_j\|^2}{s^2}\right) = exp\left(-\frac{\|x_i, x_j\|^2}{\gamma}\right) \tag{4.13}$$

where s^2 parameter is known as the width of the kernel and it is usually referred as γ.

This kernel is independent of the position of the data set with respect to the origin, it only utilizes the distances $\|x_i, x_j\|$ between points. For the Gaussian kernel no finite mapping $\Phi(x)$ of point x can be given. Because an infinite number of new points can be added (with $K(x_i, x_j) \cong 0$), the kernel space can be infinitely extended. It is shown that the data is mapped on a unit hypersphere in an infinite dimensional feature space.

The corresponding x_i and x_j are then the most dissimilar points situated at the boundary of the data set. The dissimilarity is measured with the distance s. These points will become the before mentioned *support vectors* ($\alpha_i > 0$). The number of kernels and their weights are obtained automatically by the quadratic optimization procedure [23].

The one-class SVM approach proceeds in feature space by determining the hyperplane h such that it separates the dataset from the origin with maximal margin ρ, while being as far as possible from it.

$$h \cdot x_i \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall_{i=1,\ldots,N} \tag{4.14}$$

and the function to evaluate the new test points z can be re-defined by:

$$f_{\text{SVM}}(z; h, \rho) = I(h \cdot z \leq \rho) \tag{4.15}$$

Minimizing the structural error ϵ_{struct} of the hyperplane results in the following problem:

$$\min \left(\frac{1}{2} \|h\|^2 - \rho + \frac{1}{\nu N} \sum_{i=1}^{N} \xi_i \right) \tag{4.16}$$

where ρ adjusts the fraction of data that are allowed to be on the wrong side of w (outliers that do not belong to $\mathbb{R}^d$).

The regularization parameter $\nu \in (0, 1)$ is a user defined parameter indicating the fraction of the data that should be accepted by the description. It can be compared with the parameter C in the formula (4.8).

In order to adjust the SVM for optimal results, the parameter γ (4.13) can be tuned to control the width of the kernel, that is, large values of γ lead to flat decision boundaries. Also, ν is an is an upper bound on the fraction of outliers and a lower bound on the fraction of SVs [26].

4.6. Jaccard Index

With the main objective to find an appropriate statistical representation for comparing the similarity and diversity of samples that are tested in the one-class classification process, the *Jaccard Index*, also known as the *Jaccard similarity coefficient*, is introduced.

The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A, B) = \frac{|A \bigcup B|}{|A \bigcap B|} \tag{4.17}$$

Given two sets, A and B, each with n binary attributes, the Jaccard coefficient is a useful measure of the overlap that A and B share with their attributes. Translated to

our problem, will measure the similarity between the testing sample set and the output predicted by the one-class classifier. Each attribute of A and B can either be 0 or 1.

There are four possible combinations of attributes (see Fig 4.4). First one, when both have value of 1, is named **T**rue **P**ositive (TP). Second, when attribute of A is 0 and attribute of B is 1, is considered a **F**alse **P**ositive (FP). Third, when attribute of A is 1 and attribute of B is 0, a **F**alse **N**egative (FN). And last one, when both attributes are 0, a **T**rue **N**egative (TN).

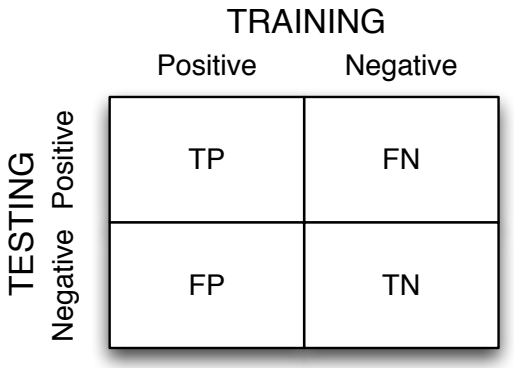


Figure 4.4.: Explanation of the similarity between the testing samples and the output of the one-class classifier.

The Jaccard similarity coefficient, J, is give as:

$$J = \frac{TP}{FN + FP + TP} \tag{4.18}$$

A J value of 1 will represent an optimal performance of the classifier. This will be obtained when the rate of False Negative and False Positive values is zero, that means, when it is no error in the classification test.

5. Implementation

5.1. Acoustic Signal Recordings

The first step of all the classification and rejection process is to record sounds of different appliances. All the training and testing process is based on this data, so this step should be done carefully.

5.1.1. Setup

An array of microphones which is placed on the head of the robot is used for the recordings. The microphones are distributed as follows: Two of them are placed on the positions of the human's ears, two in the front, and two more are located on the back of the robot's head. The distance between the two ear microphones is 19 cm, between the front and rear microphones 23 cm, front microphones have a distance of 6 cm and 4.5 cm between both on the back.

Figure 5.1.: Robot head with the microphones

Figure 5.2.: MCE 60 microphone dimensions

The array consists of six Beyerdynamic MCE 60 lapel microphones. These condenser microphones are based on electrect technology, a permanently-charged dielectric material,

so they don´t need a polarizing power supply to work. Their polar pattern appears omnidirectional and, as shown in figure (5.3), the frequency response is very linear between 20-8000 Hz [27].

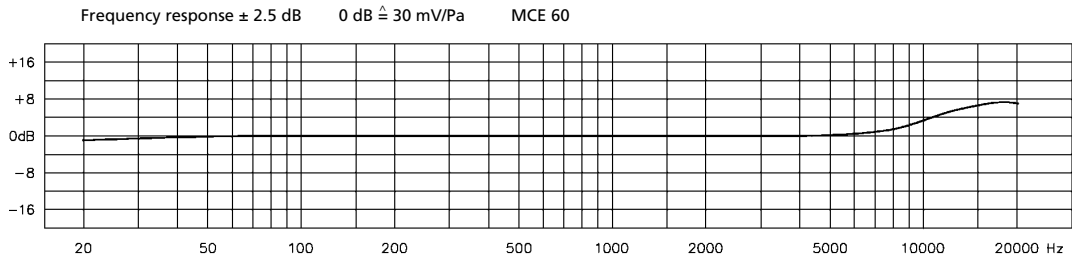


Figure 5.3.: Frequency response curve (measuring tolerance ± 2.5 dB) for the MCE 60 microphones. Extracted from [27]

For the AD/DA conversion and for the pre-amplification of the microphone signals, a MOTU *8pre* firewire audio interface is used [28].

5.1.2. Experimental Procedure

When performing the recordings, the location of the appliances was set similar in both rooms: all objects were above a common office desk and a approximate distance of 1,5 m from the robot head. The initial set of appliances consist of:

- **B**lender (B)
- **B**read **C**utter (BC)
- **C**offee **M**achine (CM)
- **M**ixer (M)
- **T**ele**p**hone (TP)
- **T**oaster (T)

Figure 5.4.: Set of the used kitchen appliances

The Coffee Machine has different phases in his process. This phases have different sounds, so, this appliance separated into four sub-states: Brewing (CM Br), Disposing (CM Di), Grinding (CM Gr) and pressing (CM Pr). A general model plus a specific model for each state will be created.

As will be explained later, for the verification of the rejection process, two new sound sources were recorded afterwards:

- **A**larm **C**lock (AC)
- **W**ater **H**eater (WH)

Figure 5.5.: New appliances

The sound sources have been recorded using a sample frequency of 16 KHz and stored for further processing using Matlab. [29].

The next step of the processing is to divide each recording into frames so that each frame will be converted into one of the parametric representations (see Sec. 3.1). Each feature vector is created from a frame with a length of 455 samples, which is approximately 28 ms of recording, what will give us 35 feature vectors per second.

In order to exclude segments with no information, a sound activity detection based on normalized energy is utilized in the baseline system. Thus, experiments are tested with three different settings of **S**ignal to **N**oise **R**atio:

- 3 dB
- 4,5 dB
- 6 dB

5.2. Multi class classifier

As mentioned before, OPASCA system has already implemented a multi class classification stage for recognition and identification of persons and different kitchen appliances. As explained in [14], OPASCA system is using MFCC features (see Sec. 3.1.1) and a GMM (see Sec .3.3.1) for this task. MFCC, in combination with GMM, have become the most used analysis method for automatic text-independent speech and speaker recognition [6].

The sound signal of each appliance is characterized by vectors of 12 MFCC, which represent the acoustical features of each object. The implementation of the GMM is using 40 mixtures of gaussians for the classification of an object. It is also using 40 mixtures for building the general model of the 'person' class. This class is used for making a first approach in the person classification. It just gives the decision between the input sound source is an object or a person.

When something is considered by the first GMM as a 'person' class, a more exact classification is done by using the UBM (see Sec. 3.3.3). The trained model in the OPASCA system is using 512 Gaussian mixtures for this classification stage. Data extracted from approximately three hours of speech was used for approximating a model for the persons speech [14].

5.3. One-class SVM

Between the many toolboxes and implementations available, a implementation from the ASI (Architecture des Systemes d'information), Department of the National Institute of Applied Sciences (INSA) in Rouen (France) is used [30]. The toolbox provides training and testing algorithms for one-class Support Vector Machine. With this toolbox, we are able to choose between different kernel functions, adjust the width of the kernel and define the number of support vectors that want's to be used. The implementation is fully done in Matlab, so can be easily adapted to the OPASCA system.

As the possibility of adding or erasing new appliances to the data set has to be taken into account, the one-class classification stage is formed by many one-class Support Vector machines. With the training data of each appliance an specific model for each one is created. By solving the problem with a modular structure, we will allow the robot to learn easily new objects. If only a unique classifier is created, all the data of all the known objects has to be used to estimate the model. Modifications in this big model will take more time. Furthermore, the estimation of the boundary around this data will become complex. Many of the objects could overlap and it would become more difficult to fit correctly the boundary. If the model is not correctly fitted, new input objects could lie in areas between this objects and be considered known when they are not.

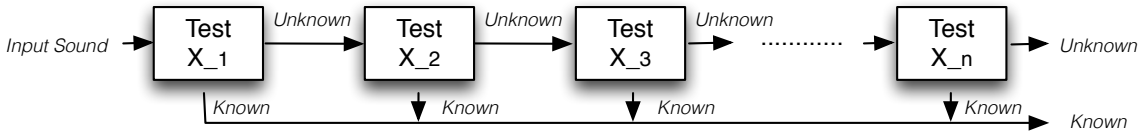


Figure 5.6.: Cascade process for a n number of model tests.

When computational power becomes a problem, as theoretically only one model will correspond to the tested object, the system could be simplified by skipping the testing phase for the other models when we get a positive or known sound source.

In addition, to measure the effectiveness of the algorithm in different environments, all the test are done for two certain cases: First one, cross-validation, where the data used for training and testing is recorded in the same office room. As only one recording of around 5 minutes for each appliance in the same room is available, two thirds of the data are used for training and the last third for testing. Second one, mismatched conditions, the training of the model and testing of the objects is done in different rooms. If the model is created in room A, the objects are tested with data from the room B, and vice-versa.

5.3.1. ν and γ parameters estimation

The correct parameter election of the SVM is one of the most important aspects for a good performance. *Gamma*, the width of the kernel, and *nu*, the number of support vectors and errors, have to be defined in this case. When a good combination of these two parameters is set correctly, good performances will be achieved, but when they are set incorrectly, the method might completely fail.

Even if each object has his own model, all the models has to be calculated using the same values of *nu* and *gamma*. When a new object want's to be added to the set of known sources, the parameters for adding this new appliance have also to be the same. Thus, a combination flexible enough to fit all different object, independent of the given features and the performance environment, has to be found.

One of the common methods for estimating this parameters is to perform a grid search in a cross-validation process [31]. With the same data for training the models and testing, a iteration process, where the combinations of parameters is changed, is performed. Then, the highest accuracy results are saved and the best combination presented.

A big problem appears when performing this grid search: As the test is done in cross-validation conditions, the combination of parameters obtained doesn't fit the boundary good enough for the cases of known objects, that means, the models appear to be under-fitted. Rates of 95% of accuracy were reached in the grid search, but as the models are under-fitted, this is, they do not fit correctly the training data, any unknown object is able to be rejected in the posterior test (high rate of false positives (FP)).

A good balance for the model has to be found. If the parameters are creating a very tied model, when unexpected changes as presence of noise, big changes in the environment or spectrally very different sounds appear, the accuracy of the test will decrease dramatically. On the other hand, if the model is too flexible, not only the data that corresponds will fit in the model, but also many other outliers will be considered target points, by not allowing the rejection. A proper balance between over-fitting and under-fitting has to be achieved.

Therefore, a time consuming try and error manual process is performed for finding the the best combinations of parameters. For the verification of this parameter election, original values for parametric representation (12 MFCC features) and energy normalization (3 dB of Signal to Noise Ratio) of the OPASCA system were chosen.

After the exhaustive search of the parameters, the chosen values are:

- γ: 0,55
- ν: 0,05

With this combination, 0,821 and 0,698 of accuracy (measured with the Jaccard Index (see Sec. 4.6)) are obtained for cross-validation and for mismatched conditions, respectively.

As the last step to prove the best combination, it is important to verify that the values will also work with different appliances and different environments . For proving the flexibility to different environments, this estimation is performed with a different data set than the one used for obtaining the general results. Also, as mentioned before, for proving the independency to specific features, two new appliances, the Alarm Clock (AC) and the Water Heater (WH) were added.

6. Evaluation

The following subsections disseminate the obtained different results. More detailed and extended results are included in the Appendices A and B. Results for combinations of different parametric representations and Signal-to-Noise Ratio (SNR) are presented.

6.1. Parametric representation combination tests

A wide range of possibilities exist for parametrically representing the sound sources for the object rejection task. MFCC is the best known and most popular [32] and it was already used in the OPASCA system. In order to obtain the best results, different parametric representations as LPC (see Sec. 3.1.2) and LPCC (see Sec.3.1.3) are also tested. To prove how each representation responds in the SVM, all of them are tested independently with different numbers of features:

- 12 MFCC, 16 MFCC
- 6 LPC, 12 LPC
- 6 LPCC, 12 LPCC

The possibility of combinations between different representations is taken into account. This combinations are based on MFCC features as the main describer of the acoustic features of the sound sources:

- 12 MFCC + 6 LPC
- 12 MFCC + 12 LPC
- 16 MFCC + 6 LPC , 16 MFCC + 12 LPC
- 12 MFCC + 6 LPCC, 12 MFCC + 12 LPCC
- 16 MFCC + 6 LPCC, 16 MFCC + 12 LPCC

Parametric representation \ SNR		3 dB	4,5 dB	6 dB
MFCC 12	*CV*	0,821	0,821	0,800
	MM	0,689	0,674	0,689
MFCC 16	*CV*	**0,857**	0,857	**0,882**
	MM	**0,811**	0,789	**0,811**
LPC 6	*CV*	0,372	0,360	0,561
	MM	0,330	0,333	0,319
LPC 12	*CV*	0,516	0,533	0,533
	MM	0,431	0,413	0,449
LPCC 6	*CV*	0,400	0,410	0,410
	MM	0,392	0,387	0,387
LPCC 12	*CV*	0,552	0,552	0,582
	MM	0,403	0,446	0,453

Figure 6.1.:
Different parametric representation tests for cross-validation (CV) and mismatched conditions (MM) using different Signal-to-Noise Ratios.

Parametric representation \ SNR		3 dB	4,5 dB	6 dB
MFCC 12 + LPC 6	*CV*	0,889	0,889	0,865
	MM	0,707	0,700	0,700
MFCC 12 + LPC 12	*CV*	0,889	0,865	0,889
	MM	0,725	0,744	0,744
MFCC 16 + LPC 6	*CV*	0,938	**0,938**	0,938
	MM	0,794	**0,824**	0,788
MFCC 16 + LPC 12	*CV*	0,938	0,938	**0,938**
	MM	0,818	0,818	**0,818**
MFCC 12 + LPCC 6	*CV*	0,941	0,889	0,889
	MM	0,730	0,730	0,806
MFCC 12 + LPCC 12	*CV*	0,941	0,914	0,941
	MM	0,722	0,750	0,743
MFCC 16 + LPCC 6	*CV*	0,938	0,938	**0,938**
	MM	0,735	0,758	**0,788**
MFCC 16 + LPCC 12	*CV*	0,938	0,906	0,906
	MM	0,758	0,758	0,788

Figure 6.2.:
Combinations of parametric representations tests for cross-validation (CV) and mismatched conditions (MM) using different Signal-to-Noise Ratios.

Figure (6.1) shows the single parametric representation tests for MFCC, LPC and LPCC. The highest accuracy is given by MFCC. The LPC and LPCC for both 6 and 12 number of features show that the performance is not acceptable. For single parametric representation, the best results are obtained with 12 and 16 MFCC features. The highest accuracy is obtained with 16 MFCC and 6 dB of SNR.

Figure (6.2) shows the combinations of parametric representations. 16 MFCC in combination with 6 LPC, 12 LPC or 6 LPCC appears to be the best choice. Between the three mentioned cases, the highest accuracy is obtained with 16 MFCC + 6 LPC and a SNR of 4,5 dB.

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	86,59	7,24	0,50	5,90	4,21	1,31	0,34	2,07	3,04	0,00	8,10	0,46
B	1,01	80,08	1,00	1,97	1,80	0,08	0,00	0,23	4,30	0,00	3,01	0,01
BC	0,06	1,26	75,44	25,78	22,18	1,74	0,13	3,65	0,11	0,00	4,33	0,01
CM	0,11	0,43	0,74	82,59	51,14	21,95	20,18	21,69	0,04	0,00	6,15	7,60
CM Br	0,20	1,26	2,27	88,31	89,33	13,18	7,58	42,78	0,23	0,00	14,92	17,29
CM Di	0,14	0,31	0,68	70,18	49,34	82,03	18,19	17,63	0,87	0,00	5,66	1,40
CM Gr	0,17	0,13	0,03	81,23	35,93	16,95	87,46	8,69	0,00	0,00	4,69	22,82
CM Pr	3,21	0,47	3,92	76,07	57,16	9,76	6,31	74,40	0,12	0,00	24,60	8,35
M	0,41	2,84	0,19	0,48	0,51	0,27	0,00	0,12	42,35	0,00	4,61	0,03
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	93,09	0,03	0,00
T	5,38	3,78	1,25	19,46	12,62	5,99	1,54	7,26	20,90	0,00	87,27	2,09
WH	0,03	0,06	0,01	40,00	25,02	4,11	15,98	5,16	0,00	0,00	3,78	88,14

Figure 6.3.: Mean Cross-Validation (CV) results for 16 MFCC + 6 LPC (4,5 dB SNR).

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	56,89	3,24	0,96	4,90	4,16	1,62	0,30	4,69	1,44	0,00	11,36	0,43
B	1,52	62,73	1,37	3,48	3,16	0,71	0,05	0,84	15,17	0,00	5,15	0,13
BC	0,40	2,73	81,32	25,05	19,58	4,90	0,21	2,71	0,98	0,00	6,82	0,05
CM	0,60	1,55	1,65	72,37	44,44	12,72	7,86	20,78	0,96	0,00	10,95	11,40
CM Br	2,45	5,10	4,27	76,76	77,92	14,19	5,69	39,37	4,14	0,00	23,53	21,64
CM Di	0,27	0,65	0,62	45,48	25,32	55,26	6,74	9,84	1,31	0,00	6,59	1,24
CM Gr	0,25	0,20	0,07	58,31	20,84	10,08	70,36	5,96	0,16	0,00	8,12	28,06
CM Pr	3,89	1,33	2,61	62,65	43,75	10,59	5,84	53,72	1,97	0,00	27,68	8,83
M	0,10	1,38	0,10	0,50	0,47	0,38	0,02	0,14	68,96	0,00	5,50	0,05
TP	0,01	0,02	0,01	0,05	0,04	0,00	0,00	0,02	0,01	34,49	0,04	0,00
T	7,67	5,46	1,48	20,84	15,85	11,95	2,44	11,24	33,72	0,00	77,53	4,88
WH	0,68	1,30	0,05	46,71	28,51	5,88	24,40	12,85	1,03	0,00	14,69	85,60

Figure 6.4.: Mean Mismatched Condition (MM) results for 16 MFCC + 6 LPC (4,5 dB SNR).

For this best combination, results for Cross-Validation (CV) and Mismatched Conditions (MM) are shown. The results representing the mean between the test results of two rooms.

In the two cross-validation (CV) cases and two Mismatched Condition (MM) the results are meaned. For the complete results see Appendix A.

In figure (6.3) the results for the cross-validation (CV) case are shown. All the objects are rejected correctly. The only wrong value can be found for the mixer. As explained before, cross-validation is performed with the first two thirds of the data for training and the last third for testing. As it is not possible to set the mixer on a table, the recordings of the mixer were performed by holding the appliance with both hands. Changes between the beginning and the ending of the recordings could justify this accuracy decrease.

For showing how the rejection process will work, we assume that one model of the objects is erased. We will assume cross-validations conditions and we will take the Alarm Clock (AC) as an unknown object (first column of the tables should be omitted, the model is not created). For verifying that it is an unknown object, as explained in image (5.6), every model has to be tested. After obtaining al the results, we can see that the biggest value corresponds the telephone (8,10 %). As the value is under 50 %, it is labeled as unknown, that is, it is rejected.

In the case the case the model of the Alarm Clock (AC) is added, the highest result corresponds to the test with it's model (86,59 %). As the value is over 50 % the object will be labeled as known.

6.2. Signal acquisition duration tests

For the obtained best combination, 16 MFCC + 6 LPC (4,5 dB SNR), a real time response of the classifier is tested. For this, results for 0.5, 1, 2, 5 and 10 seconds of testing data acquisition are shown. Further results of each specific period of time are shown in appendix B.

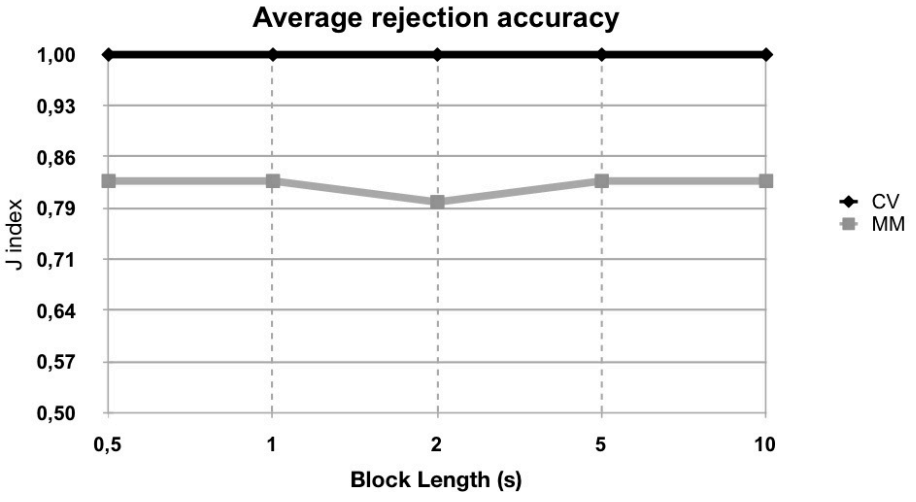


Figure 6.5.: Average rejection accuracy for specific period of time training data.

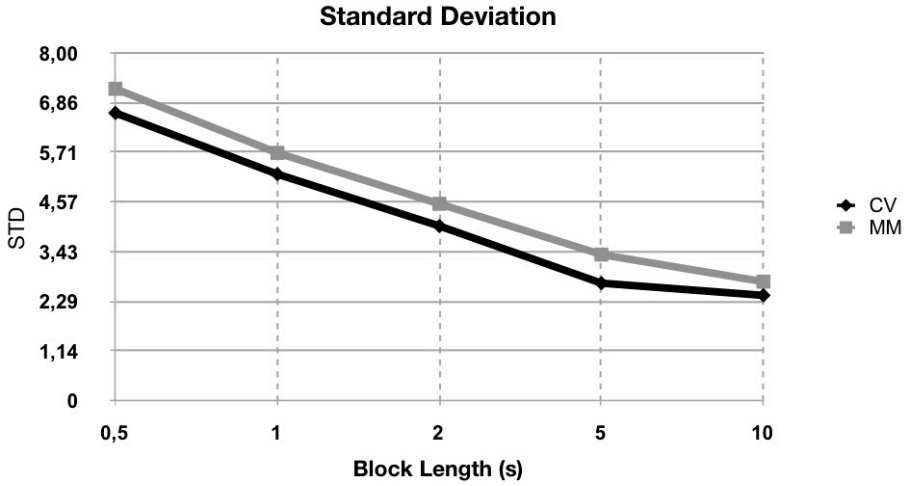


Figure 6.6.: Standard deviation of average rejection accuracy for specific period of time training data.

The accuracy values appear to be quite constant over time. However, the standard deviation decreases having more testing data. This corresponds to a stabilization of the tests over time. This is, when more testing data is available, the results will become stable.

7. Conclusions and Outlook

This thesis has presented the design and development of a sound source rejection approach. Acoustically observable kitchen appliances can be labeled as known or unknown.

The one-class Support Vector Machine shows to be capable of performing the desired rejection. This algorithm is intended for been used in the OPASCA system. The combination of MFCC and LPC as parametric representations shows to be the best choice for this task. More precisely, a Signal-to-Noise Ration of 4,5 dB and 16 MFCC features in combination with 6 LPC features is the best choice.

The experimental results present that the accuracy of Cross-Validation (CV) is considerably high. When the training and the testing is done in the same environment, the sound source rejection is performed quite correctly.

Under mismatched conditions (MM) the performance of the system decreases. Even if the unknown objects are still able to be rejected, some models appear over-fitted, and known objects are labeled as unknown. The bad case of unknown objects been labeled as known it is performed correctly. The accuracy obtained is lower than in the cross-validation case.

Acquisition of training data in short periods of time shows correct results. When more training data is acquired, the standard deviation decreases, which means, the classifier becomes more stable.

Future work could be performed in improving the results for mismatched condition. For example, a standard room model can be introduced in order to avoid the environment dependency.

As in every environment background noises are appearing, for doing the simulations more realistic, simulated background noise could be added to the signals.

In addition, to prove the robustness of the implemented classifiers, more kitchen appliances could be added. Finally, the ability to reject unknown sound sources opens the way of learning new objects in the environment of the robot.

Unsupervised creation of new models will allow the humanoid robot to adapt faster to the changing scenarios.

A. Appendix

Cross-validation (CV) and Mismatch conditions (MM) extended results for the combination of 16 MFCC + 6 LPC and a SNR of 4,5 dB.

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	86,69	0,89	0,36	3,60	2,46	1,53	0,18	2,17	0,18	0,00	4,88	0,89
B	0,03	78,66	0,84	1,60	1,34	0,11	0,00	0,20	7,31	0,00	1,76	0,03
BC	0,00	0,17	71,70	26,68	24,09	3,15	0,20	6,05	0,00	0,00	6,45	0,03
CM	0,18	0,41	1,35	82,74	50,66	23,12	19,70	21,55	0,03	0,00	7,10	10,64
CM Br	0,27	0,82	4,03	86,72	89,23	16,72	6,68	40,88	0,32	0,00	14,23	16,80
CM Di	0,00	0,13	1,22	75,30	52,02	84,45	19,39	11,13	0,77	0,00	6,46	2,24
CM Gr	0,27	0,07	0,00	80,99	38,76	15,82	90,77	8,69	0,00	0,00	4,01	35,03
CM Pr	4,33	0,12	7,03	79,16	59,13	12,88	8,67	82,79	0,00	0,00	34,07	9,25
M	0,00	0,67	0,00	0,17	0,22	0,28	0,00	0,00	43,61	0,00	3,51	0,03
TP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	96,10	0,00	0,00
T	5,45	3,98	1,66	16,60	12,81	7,50	0,95	6,59	22,60	0,00	91,10	1,53
WH	0,03	0,09	0,03	37,72	25,58	7,40	17,24	1,81	0,00	0,00	2,94	87,85

Figure A.1.: Cross-Validation (CV) results (Room A) for 16 MFCC + 6 LPC (4,5 dB SNR).

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	86,50	13,59	0,64	8,20	5,97	1,08	0,51	1,98	5,90	0,00	11,33	0,03
B	1,99	81,50	1,15	2,33	2,25	0,06	0,00	0,26	1,30	0,00	4,27	0,00
BC	0,11	2,35	79,18	24,87	20,26	0,34	0,06	1,26	0,22	0,00	2,21	0,00
CM	0,04	0,46	0,13	82,43	51,63	20,79	20,66	21,82	0,04	0,00	5,20	4,56
CM Br	0,12	1,69	0,52	89,89	89,44	9,63	8,48	44,67	0,14	0,00	15,61	17,78
CM Di	0,28	0,49	0,14	65,05	46,67	79,61	16,99	24,13	0,97	0,00	4,85	0,55
CM Gr	0,06	0,19	0,06	81,47	33,10	18,08	84,15	8,69	0,00	0,00	5,37	10,61
CM Pr	2,10	0,81	0,81	72,99	55,18	6,64	3,96	66,01	0,23	0,00	15,13	7,45
M	0,83	5,00	0,38	0,80	0,80	0,27	0,00	0,24	41,10	0,00	5,72	0,03
TP	0,03	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,03	90,08	0,06	0,00
T	5,32	3,58	0,84	22,33	12,44	4,48	2,13	7,93	19,20	0,00	83,44	2,64
WH	0,04	0,04	0,00	42,27	24,47	0,82	14,71	8,51	0,00	0,00	4,63	88,43

Figure A.2.: Cross-Validation (CV) results (Room B) for 16 MFCC + 6 LPC (4,5 dB SNR).

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	58,85	3,52	1,80	7,59	7,03	2,44	0,37	7,32	0,28	0,00	19,28	0,81
B	1,61	40,53	2,03	5,73	5,17	1,29	0,07	1,58	1,56	0,00	7,26	0,25
BC	0,07	1,29	86,55	30,79	24,58	9,36	0,31	4,12	0,00	0,00	4,89	0,08
CM	0,05	0,06	0,48	74,05	45,12	20,46	8,08	16,31	0,03	0,00	2,42	17,79
CM Br	0,12	0,30	1,26	77,52	78,20	22,77	6,86	28,51	0,24	0,00	4,48	32,53
CM Di	0,19	0,07	0,23	32,35	22,55	58,58	2,78	4,44	0,79	0,00	3,31	1,99
CM Gr	0,11	0,04	0,06	46,92	17,92	13,51	63,12	3,37	0,00	0,00	1,92	41,96
CM Pr	0,47	0,00	1,79	65,72	46,64	14,57	7,93	50,87	0,00	0,00	7,31	15,24
M	0,11	2,44	0,17	0,94	0,86	0,59	0,02	0,27	60,53	0,00	7,73	0,10
TP	0,01	0,03	0,01	0,10	0,08	0,00	0,00	0,03	0,01	18,92	0,06	0,00
T	2,33	1,88	2,27	33,12	25,89	16,03	3,78	17,78	17,17	0,00	72,86	9,64
WH	0,03	0,00	0,00	30,56	16,77	2,33	10,05	4,26	0,00	0,00	0,89	88,00

Figure A.3.: Mismatched Condition (MM) results (Training A- Testing B) for 16 MFCC + 6 LPC (4,5 dB SNR).

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	54,92	2,96	0,13	2,21	1,29	0,81	0,23	2,05	2,60	0,00	3,43	0,05
B	1,43	84,92	0,71	1,23	1,15	0,13	0,02	0,10	28,78	0,00	3,03	0,00
BC	0,73	4,16	76,09	19,30	14,57	0,43	0,10	1,29	1,95	0,00	8,74	0,01
CM	1,14	3,04	2,82	70,69	43,75	4,97	7,64	25,24	1,88	0,00	19,47	5,01
CM Br	4,77	9,90	7,27	76,00	77,63	5,61	4,51	50,22	8,04	0,00	42,57	10,74
CM Di	0,36	1,24	1,00	58,61	28,09	51,93	10,71	15,25	1,83	0,00	9,87	0,49
CM Gr	0,39	0,36	0,07	69,69	23,76	6,66	77,60	8,56	0,32	0,00	14,31	14,16
CM Pr	7,30	2,66	3,44	59,57	40,86	6,60	3,75	56,56	3,95	0,00	48,05	2,42
M	0,08	0,32	0,02	0,05	0,07	0,17	0,01	0,01	77,39	0,00	3,26	0,00
TP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	50,07	0,02	0,00
T	13,00	9,03	0,70	8,57	5,82	7,86	1,10	4,70	50,28	0,00	82,20	0,11
WH	1,34	2,60	0,10	62,85	40,25	9,43	38,75	21,43	2,05	0,00	28,50	83,19

Figure A.4.: Mismatched Condition (MM) results (Training B- Testing A) for 16 MFCC + 6 LPC (4,5 dB SNR).

B. Appendix

Cross-validation (CV) results after acquisition of 0.5, 1, 2, 5 and 10 seconds of testing data.

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,79	7,49	0,72	6,06	4,51	1,69	0,43	3,96	4,60	0,00	9,07	0,45
B	1,32	90,17	0,90	1,92	1,83	0,17	0,00	0,38	6,63	0,00	3,22	0,01
BC	0,16	1,31	90,82	25,64	23,37	2,29	0,16	4,16	0,73	0,00	4,59	0,03
CM	0,18	0,44	0,92	82,54	51,82	25,55	22,54	26,67	0,13	0,00	7,28	8,16
CM Br	0,34	1,53	2,89	88,34	89,83	16,64	9,39	52,35	0,62	0,00	16,82	18,01
CM Di	0,26	0,45	0,86	69,92	49,13	87,34	18,84	20,10	1,34	0,00	6,26	1,55
CM Gr	0,22	0,09	0,06	81,17	36,07	19,00	92,29	11,26	0,09	0,00	6,11	26,33
CM Pr	3,39	0,50	5,00	76,00	58,06	12,39	7,89	85,39	0,44	0,00	30,44	8,67
M	0,32	2,74	0,18	0,44	0,51	0,41	0,00	0,20	83,42	0,00	4,49	0,03
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	94,97	0,03	0,00
T	5,64	3,65	1,23	19,09	13,11	10,35	1,73	10,01	28,84	0,00	91,88	1,85
WH	0,07	0,05	0,01	39,81	24,85	4,97	19,11	7,34	0,02	0,00	4,87	93,85

Figure B.1.: Cross-Validation (CV) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 0,5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	7,87	8,17	2,33	8,37	6,64	3,65	1,58	5,99	5,63	0,00	10,91	1,61
B	3,31	12,09	2,55	5,42	5,55	0,99	0,00	2,17	10,23	0,00	6,99	0,19
BC	0,65	2,59	9,27	23,43	21,44	4,71	0,93	6,52	1,70	0,00	7,30	0,38
CM	0,94	1,55	2,19	11,13	15,09	11,78	11,33	14,12	0,85	0,00	6,89	6,98
CM Br	1,47	3,47	6,31	14,71	12,27	18,88	13,01	24,85	2,13	0,00	15,66	18,24
CM Di	1,13	1,58	2,24	18,42	22,90	9,38	26,06	15,78	3,52	0,00	9,57	3,93
CM Gr	1,08	0,50	0,59	13,70	19,18	24,10	8,88	18,45	0,50	0,00	8,77	13,44
CM Pr	4,69	1,56	7,85	16,70	15,87	11,21	8,05	10,15	1,03	0,00	13,66	8,32
M	1,59	6,44	1,28	2,56	2,80	2,45	0,00	1,36	15,05	0,00	12,62	0,39
TP	0,19	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,19	7,89	0,27	0,00
T	7,25	5,83	3,08	15,91	11,66	8,98	3,61	10,36	19,15	0,00	8,02	4,18
WH	0,62	0,50	0,20	17,34	15,40	5,76	12,40	7,44	0,23	0,00	6,23	6,01

Figure B.2.: Cross-Validation (CV) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 0,5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,74	7,36	0,69	5,85	4,39	1,65	0,40	3,83	4,61	0,00	8,84	0,43
B	1,32	90,19	0,93	1,92	1,84	0,18	0,00	0,37	6,62	0,00	3,23	0,01
BC	0,15	1,32	90,86	25,53	23,27	2,31	0,16	4,11	0,73	0,00	4,60	0,03
CM	0,19	0,44	0,93	82,52	51,76	25,60	22,58	26,68	0,13	0,00	7,29	8,16
CM Br	0,35	1,53	2,89	88,28	89,79	16,61	9,43	52,10	0,63	0,00	16,81	18,04
CM Di	0,27	0,47	0,86	69,87	49,10	87,44	18,84	20,12	1,36	0,00	6,15	1,54
CM Gr	0,22	0,09	0,06	81,06	36,04	18,96	92,23	11,29	0,09	0,00	6,03	26,36
CM Pr	3,31	0,51	5,14	76,00	57,71	12,29	7,71	85,20	0,40	0,00	30,23	8,40
M	0,33	2,74	0,19	0,46	0,51	0,36	0,00	0,20	83,62	0,00	4,40	0,03
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	95,07	0,03	0,00
T	5,75	3,74	1,23	19,15	13,23	10,31	1,73	10,08	28,77	0,00	91,83	1,84
WH	0,08	0,05	0,01	40,09	24,98	5,00	19,30	7,30	0,02	0,00	4,83	93,86

Figure B.3.: Cross-Validation (CV) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 1 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	6,35	4,74	1,60	5,55	4,31	2,58	1,06	4,04	3,35	0,00	6,96	1,17
B	2,57	10,49	2,01	4,12	4,40	0,78	0,00	1,57	8,39	0,00	5,52	0,14
BC	0,48	1,91	7,55	19,70	18,14	3,88	0,67	5,27	1,24	0,00	5,89	0,28
CM	0,67	1,09	1,63	8,79	11,83	8,95	8,77	11,52	0,65	0,00	5,15	5,20
CM Br	1,06	2,69	5,73	12,73	10,38	16,00	11,55	22,89	1,62	0,00	13,99	16,39
CM Di	0,80	1,23	1,65	14,93	20,51	7,11	22,20	13,06	2,42	0,00	7,78	2,93
CM Gr	0,78	0,36	0,43	11,65	16,02	21,40	7,64	16,35	0,36	0,00	6,94	10,39
CM Pr	2,93	1,08	6,18	10,10	9,91	7,23	5,92	7,03	0,65	0,00	10,14	6,10
M	1,29	5,11	0,93	2,02	2,21	1,68	0,00	0,98	12,59	0,00	10,41	0,28
TP	0,14	0,14	0,00	0,00	0,00	0,00	0,00	0,00	0,14	5,88	0,20	0,00
T	5,12	4,09	2,16	10,59	7,68	6,01	2,49	7,17	12,08	0,00	5,66	2,92
WH	0,45	0,37	0,14	14,97	13,57	4,65	10,51	6,08	0,16	0,00	4,81	4,49

Figure B.4.: Cross-Validation (CV) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 1 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,76	7,49	0,74	5,93	4,44	1,63	0,42	3,88	4,66	0,00	8,94	0,42
B	1,20	90,66	0,89	1,74	1,65	0,17	0,00	0,29	6,52	0,00	3,10	0,01
BC	0,14	1,30	91,08	25,91	23,64	2,30	0,17	4,20	0,74	0,00	4,65	0,03
CM	0,18	0,44	0,93	82,59	51,79	25,59	22,64	26,64	0,12	0,00	7,25	8,18
CM Br	0,33	1,53	2,93	88,21	89,70	16,59	9,46	51,96	0,64	0,00	16,69	18,00
CM Di	0,28	0,42	0,89	70,18	49,69	87,56	18,79	20,31	1,34	0,00	6,31	1,55
CM Gr	0,23	0,10	0,07	81,15	36,03	18,75	92,41	10,99	0,10	0,00	6,05	26,55
CM Pr	3,21	0,54	5,12	75,77	57,14	12,62	7,98	85,06	0,42	0,00	29,40	8,69
M	0,33	2,72	0,18	0,43	0,47	0,30	0,00	0,19	83,96	0,00	4,20	0,01
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	95,12	0,03	0,00
T	5,62	3,75	1,22	19,25	13,18	10,28	1,72	10,16	28,65	0,00	91,87	1,82
WH	0,08	0,05	0,01	40,15	24,85	4,92	19,30	7,28	0,02	0,00	4,77	93,88

Figure B.5.: Cross-Validation (CV) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 2 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	4,98	3,87	1,29	4,67	3,40	2,04	0,82	2,99	2,71	0,00	5,92	0,83
B	1,80	7,62	1,63	3,05	2,87	0,53	0,00	0,71	6,85	0,00	4,04	0,10
BC	0,29	1,50	5,37	12,97	11,90	2,93	0,46	3,73	1,02	0,00	4,00	0,20
CM	0,45	0,80	1,25	6,77	9,31	7,04	7,00	9,38	0,43	0,00	3,93	3,97
CM Br	0,76	2,18	5,38	11,27	8,87	13,89	10,21	21,06	1,40	0,00	12,44	14,50
CM Di	0,62	0,80	1,19	13,18	16,90	5,46	17,85	9,44	1,67	0,00	6,23	2,22
CM Gr	0,54	0,25	0,31	9,70	11,24	15,20	5,76	11,74	0,25	0,00	4,70	8,68
CM Pr	2,06	0,67	5,34	8,60	6,67	5,69	4,37	5,99	0,48	0,00	8,42	4,49
M	0,85	4,08	0,52	1,25	1,33	1,12	0,00	0,58	10,74	0,00	7,95	0,10
TP	0,10	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,10	4,61	0,14	0,00
T	3,28	2,69	1,53	5,28	4,53	3,64	1,67	4,36	6,33	0,00	4,05	1,78
WH	0,34	0,26	0,10	13,26	11,57	3,53	8,41	4,95	0,12	0,00	3,79	3,45

Figure B.6.: Cross-Validation (CV) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 2 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,68	7,66	0,73	6,04	4,53	1,70	0,44	3,93	4,75	0,00	9,06	0,42
B	1,21	90,84	0,97	1,73	1,62	0,17	0,00	0,30	6,66	0,00	3,13	0,01
BC	0,14	1,31	91,24	25,81	23,51	2,27	0,17	4,15	0,72	0,00	4,63	0,03
CM	0,18	0,44	0,93	82,51	51,68	25,67	22,65	26,53	0,12	0,00	7,24	8,16
CM Br	0,35	1,49	2,89	88,17	89,64	16,58	9,34	51,72	0,64	0,00	16,65	17,71
CM Di	0,28	0,43	0,82	70,10	49,64	87,57	19,28	20,24	1,24	0,00	6,00	1,49
CM Gr	0,25	0,11	0,07	81,57	36,93	19,67	92,37	11,61	0,11	0,00	6,29	26,81
CM Pr	3,20	0,57	5,68	76,63	57,81	13,57	8,31	85,58	0,36	0,00	31,11	9,02
M	0,31	2,72	0,18	0,41	0,47	0,30	0,00	0,18	83,96	0,00	4,19	0,01
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	95,08	0,03	0,00
T	5,46	3,76	1,14	19,07	13,05	10,28	1,74	9,93	28,74	0,00	91,81	1,79
WH	0,07	0,05	0,01	40,23	25,00	5,03	19,34	7,29	0,02	0,00	4,83	93,90

Figure B.7.: Cross-Validation (CV) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	3,30	3,47	0,75	3,84	2,81	1,66	0,59	2,30	2,13	0,00	5,05	0,62
B	1,09	6,44	1,00	1,88	1,79	0,31	0,00	0,47	4,83	0,00	2,09	0,06
BC	0,20	1,10	3,11	8,56	7,93	1,85	0,27	2,67	0,59	0,00	2,66	0,13
CM	0,31	0,55	0,91	4,72	6,38	5,63	5,00	7,14	0,29	0,00	2,54	2,94
CM Br	0,65	1,57	4,06	8,86	7,26	11,17	8,09	17,70	1,15	0,00	11,14	11,83
CM Di	0,46	0,50	0,61	5,68	8,75	2,83	8,13	4,26	0,97	0,00	1,90	1,15
CM Gr	0,30	0,21	0,20	6,79	4,97	5,89	3,63	3,62	0,21	0,00	2,53	5,38
CM Pr	1,44	0,28	4,81	5,37	4,49	3,77	3,33	2,49	0,14	0,00	4,69	2,47
M	0,58	3,04	0,43	0,93	1,04	0,72	0,00	0,44	8,60	0,00	5,52	0,06
TP	0,06	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,06	3,29	0,09	0,00
T	1,90	2,06	0,90	3,19	2,27	2,33	1,37	2,47	3,84	0,00	2,60	1,11
WH	0,18	0,17	0,07	11,92	10,39	3,15	6,87	4,15	0,08	0,00	3,06	2,43

Figure B.8.: Cross-Validation (CV) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,66	7,88	0,75	6,21	4,65	1,72	0,44	4,05	4,83	0,00	9,25	0,43
B	1,21	90,74	0,90	1,73	1,64	0,18	0,00	0,30	6,73	0,00	3,15	0,01
BC	0,14	1,31	91,12	25,81	23,51	2,30	0,17	4,19	0,72	0,00	4,60	0,03
CM	0,19	0,42	0,92	82,38	51,60	25,65	22,70	26,33	0,12	0,00	7,16	8,11
CM Br	0,34	1,52	2,99	87,77	89,30	16,64	9,23	50,79	0,66	0,00	16,47	17,17
CM Di	0,28	0,43	0,85	70,13	49,72	87,61	19,25	20,24	1,24	0,00	5,97	1,49
CM Gr	0,25	0,11	0,07	81,64	36,90	19,60	92,33	11,65	0,11	0,00	6,29	26,70
CM Pr	3,27	0,57	5,68	76,42	57,67	13,49	8,31	85,58	0,36	0,00	30,97	8,88
M	0,33	2,83	0,19	0,43	0,49	0,31	0,00	0,19	84,32	0,00	4,36	0,01
TP	0,02	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,02	94,87	0,03	0,00
T	5,47	3,85	1,19	19,26	13,12	10,33	1,74	9,94	28,66	0,00	91,76	1,79
WH	0,07	0,05	0,02	40,51	25,17	5,16	19,54	7,29	0,02	0,00	4,85	94,00

Figure B.9.: Cross-Validation (CV) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 10 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	3,13	3,46	0,75	3,71	2,69	1,59	0,56	2,22	2,07	0,00	4,92	0,46
B	1,07	5,87	0,98	1,72	1,64	0,25	0,00	0,44	4,06	0,00	2,07	0,04
BC	0,20	1,04	2,71	7,38	6,97	1,41	0,23	2,17	0,59	0,00	2,02	0,11
CM	0,24	0,48	0,81	4,42	5,95	5,39	4,44	6,73	0,25	0,00	2,34	2,62
CM Br	0,55	1,49	3,86	6,84	6,10	10,63	7,81	15,00	0,99	0,00	10,60	11,44
CM Di	0,41	0,45	0,51	4,56	8,11	2,62	8,03	3,88	0,82	0,00	1,66	0,66
CM Gr	0,22	0,21	0,17	6,47	4,47	4,09	2,94	2,55	0,21	0,00	2,29	4,12
CM Pr	1,15	0,24	4,05	4,68	3,55	3,85	3,16	1,47	0,14	0,00	3,50	2,76
M	0,58	2,83	0,43	0,86	0,92	0,62	0,00	0,44	8,61	0,00	5,11	0,04
TP	0,06	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,06	3,05	0,09	0,00
T	1,83	1,74	0,79	2,57	2,01	2,01	1,37	1,90	2,62	0,00	2,26	4,18
WH	0,15	0,14	0,05	11,52	9,51	2,31	6,50	4,01	0,08	0,00	2,85	2,19

Figure B.10.: Cross-Validation (CV) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 10 s. of testing data

Mismatch Condition (MM) results after acquisition of 0.5, 1, 2, 5 and 10 seconds of testing data.

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,79	7,49	0,72	6,06	4,51	1,69	0,43	3,96	4,60	0,00	9,07	0,45
B	1,32	90,17	0,90	1,92	1,83	0,17	0,00	0,38	6,63	0,00	3,22	0,01
BC	0,16	1,31	90,82	25,64	23,37	2,29	0,16	4,16	0,73	0,00	4,59	0,03
CM	0,18	0,44	0,92	82,54	51,82	25,55	22,54	26,67	0,13	0,00	7,28	8,16
CM Br	0,34	1,53	2,89	88,34	89,83	16,64	9,39	52,35	0,62	0,00	16,82	18,01
CM Di	0,26	0,45	0,86	69,92	49,13	87,34	18,84	20,10	1,34	0,00	6,26	1,55
CM Gr	0,22	0,09	0,06	81,17	36,07	19,00	92,29	11,26	0,09	0,00	6,11	26,33
CM Pr	3,39	0,50	5,00	76,00	58,06	12,39	7,89	85,39	0,44	0,00	30,44	8,67
M	0,32	2,74	0,18	0,44	0,51	0,41	0,00	0,20	83,42	0,00	4,49	0,03
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	94,97	0,03	0,00
T	5,64	3,65	1,23	19,09	13,11	10,35	1,73	10,01	28,84	0,00	91,88	1,85
WH	0,07	0,05	0,01	39,81	24,85	4,97	19,11	7,34	0,02	0,00	4,87	93,85

Figure B.11.: Mismatch Condition (MM) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 0,5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	7,87	8,17	2,33	8,37	6,64	3,65	1,58	5,99	5,63	0,00	10,91	1,61
B	3,31	12,09	2,55	5,42	5,55	0,99	0,00	2,17	10,23	0,00	6,99	0,19
BC	0,65	2,59	9,27	23,43	21,44	4,71	0,93	6,52	1,70	0,00	7,30	0,38
CM	0,94	1,55	2,19	11,13	15,09	11,78	11,33	14,12	0,85	0,00	6,89	6,98
CM Br	1,47	3,47	6,31	14,71	12,27	18,88	13,01	24,85	2,13	0,00	15,66	18,24
CM Di	1,13	1,58	2,24	18,42	22,90	9,38	26,06	15,78	3,52	0,00	9,57	3,93
CM Gr	1,08	0,50	0,59	13,70	19,18	24,10	8,88	18,45	0,50	0,00	8,77	13,44
CM Pr	4,69	1,56	7,85	16,70	15,87	11,21	8,05	10,15	1,03	0,00	13,66	8,32
M	1,59	6,44	1,28	2,56	2,80	2,45	0,00	1,36	15,05	0,00	12,62	0,39
TP	0,19	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,19	7,89	0,27	0,00
T	7,25	5,83	3,08	15,91	11,66	8,98	3,61	10,36	19,15	0,00	8,02	4,18
WH	0,62	0,50	0,20	17,34	15,40	5,76	12,40	7,44	0,23	0,00	6,23	6,01

Figure B.12.: Mismatch Condition (MM) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 0,5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,74	7,36	0,69	5,85	4,39	1,65	0,40	3,83	4,61	0,00	8,84	0,43
B	1,32	90,19	0,93	1,92	1,84	0,18	0,00	0,37	6,62	0,00	3,23	0,01
BC	0,15	1,32	90,86	25,53	23,27	2,31	0,16	4,11	0,73	0,00	4,60	0,03
CM	0,19	0,44	0,93	82,52	51,76	25,60	22,58	26,68	0,13	0,00	7,29	8,16
CM Br	0,35	1,53	2,89	88,28	89,79	16,61	9,43	52,10	0,63	0,00	16,81	18,04
CM Di	0,27	0,47	0,86	69,87	49,10	87,44	18,84	20,12	1,36	0,00	6,15	1,54
CM Gr	0,22	0,09	0,06	81,06	36,04	18,96	92,23	11,29	0,09	0,00	6,03	26,36
CM Pr	3,31	0,51	5,14	76,00	57,71	12,29	7,71	85,20	0,40	0,00	30,23	8,40
M	0,33	2,74	0,19	0,46	0,51	0,36	0,00	0,20	83,62	0,00	4,40	0,03
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	95,07	0,03	0,00
T	5,75	3,74	1,23	19,15	13,23	10,31	1,73	10,08	28,77	0,00	91,83	1,84
WH	0,08	0,05	0,01	40,09	24,98	5,00	19,30	7,30	0,02	0,00	4,83	93,86

Figure B.13.: Mismatch Condition (MM) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 1 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	6,35	4,74	1,60	5,55	4,31	2,58	1,06	4,04	3,35	0,00	6,96	1,17
B	2,57	10,49	2,01	4,12	4,40	0,78	0,00	1,57	8,39	0,00	5,52	0,14
BC	0,48	1,91	7,55	19,70	18,14	3,88	0,67	5,27	1,24	0,00	5,89	0,28
CM	0,67	1,09	1,63	8,79	11,83	8,95	8,77	11,52	0,65	0,00	5,15	5,20
CM Br	1,06	2,69	5,73	12,73	10,38	16,00	11,55	22,89	1,62	0,00	13,99	16,39
CM Di	0,80	1,23	1,65	14,93	20,51	7,11	22,20	13,06	2,42	0,00	7,78	2,93
CM Gr	0,78	0,36	0,43	11,65	16,02	21,40	7,64	16,35	0,36	0,00	6,94	10,39
CM Pr	2,93	1,08	6,18	10,10	9,91	7,23	5,92	7,03	0,65	0,00	10,14	6,10
M	1,29	5,11	0,93	2,02	2,21	1,68	0,00	0,98	12,59	0,00	10,41	0,28
TP	0,14	0,14	0,00	0,00	0,00	0,00	0,00	0,00	0,14	5,88	0,20	0,00
T	5,12	4,09	2,16	10,59	7,68	6,01	2,49	7,17	12,08	0,00	5,66	2,92
WH	0,45	0,37	0,14	14,97	13,57	4,65	10,51	6,08	0,16	0,00	4,81	4,49

Figure B.14.: Mismatch Condition (MM) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 1 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,76	7,49	0,74	5,93	4,44	1,63	0,42	3,88	4,66	0,00	8,94	0,42
B	1,20	90,66	0,89	1,74	1,65	0,17	0,00	0,29	6,52	0,00	3,10	0,01
BC	0,14	1,30	91,08	25,91	23,64	2,30	0,17	4,20	0,74	0,00	4,65	0,03
CM	0,18	0,44	0,93	82,59	51,79	25,59	22,64	26,64	0,12	0,00	7,25	8,18
CM Br	0,33	1,53	2,93	88,21	89,70	16,59	9,46	51,96	0,64	0,00	16,69	18,00
CM Di	0,28	0,42	0,89	70,18	49,69	87,56	18,79	20,31	1,34	0,00	6,31	1,55
CM Gr	0,23	0,10	0,07	81,15	36,03	18,75	92,41	10,99	0,10	0,00	6,05	26,55
CM Pr	3,21	0,54	5,12	75,77	57,14	12,62	7,98	85,06	0,42	0,00	29,40	8,69
M	0,33	2,72	0,18	0,43	0,47	0,30	0,00	0,19	83,96	0,00	4,20	0,01
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	95,12	0,03	0,00
T	5,62	3,75	1,22	19,25	13,18	10,28	1,72	10,16	28,65	0,00	91,87	1,82
WH	0,08	0,05	0,01	40,15	24,85	4,92	19,30	7,28	0,02	0,00	4,77	93,88

Figure B.15.: Mismatch Condition (MM) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 2 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	4,98	3,87	1,29	4,67	3,40	2,04	0,82	2,99	2,71	0,00	5,92	0,83
B	1,80	7,62	1,63	3,05	2,87	0,53	0,00	0,71	6,85	0,00	4,04	0,10
BC	0,29	1,50	5,37	12,97	11,90	2,93	0,46	3,73	1,02	0,00	4,00	0,20
CM	0,45	0,80	1,25	6,77	9,31	7,04	7,00	9,38	0,43	0,00	3,93	3,97
CM Br	0,76	2,18	5,38	11,27	8,87	13,89	10,21	21,06	1,40	0,00	12,44	14,50
CM Di	0,62	0,80	1,19	13,18	16,90	5,46	17,85	9,44	1,67	0,00	6,23	2,22
CM Gr	0,54	0,25	0,31	9,70	11,24	15,20	5,76	11,74	0,25	0,00	4,70	8,68
CM Pr	2,06	0,67	5,34	8,60	6,67	5,69	4,37	5,99	0,48	0,00	8,42	4,49
M	0,85	4,08	0,52	1,25	1,33	1,12	0,00	0,58	10,74	0,00	7,95	0,10
TP	0,10	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,10	4,61	0,14	0,00
T	3,28	2,69	1,53	5,28	4,53	3,64	1,67	4,36	6,33	0,00	4,05	1,78
WH	0,34	0,26	0,10	13,26	11,57	3,53	8,41	4,95	0,12	0,00	3,79	3,45

Figure B.16.: Mismatch Condition (MM) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 2 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,68	7,66	0,73	6,04	4,53	1,70	0,44	3,93	4,75	0,00	9,06	0,42
B	1,21	90,84	0,97	1,73	1,62	0,17	0,00	0,30	6,66	0,00	3,13	0,01
BC	0,14	1,31	91,24	25,81	23,51	2,27	0,17	4,15	0,72	0,00	4,63	0,03
CM	0,18	0,44	0,93	82,51	51,68	25,67	22,65	26,53	0,12	0,00	7,24	8,16
CM Br	0,35	1,49	2,89	88,17	89,64	16,58	9,34	51,72	0,64	0,00	16,65	17,71
CM Di	0,28	0,43	0,82	70,10	49,64	87,57	19,28	20,24	1,24	0,00	6,00	1,49
CM Gr	0,25	0,11	0,07	81,57	36,93	19,67	92,37	11,61	0,11	0,00	6,29	26,81
CM Pr	3,20	0,57	5,68	76,63	57,81	13,57	8,31	85,58	0,36	0,00	31,11	9,02
M	0,31	2,72	0,18	0,41	0,47	0,30	0,00	0,18	83,96	0,00	4,19	0,01
TP	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	95,08	0,03	0,00
T	5,46	3,76	1,14	19,07	13,05	10,28	1,74	9,93	28,74	0,00	91,81	1,79
WH	0,07	0,05	0,01	40,23	25,00	5,03	19,34	7,29	0,02	0,00	4,83	93,90

Figure B.17.: Mismatch Condition (MM) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	3,30	3,47	0,75	3,84	2,81	1,66	0,59	2,30	2,13	0,00	5,05	0,62
B	1,09	6,44	1,00	1,88	1,79	0,31	0,00	0,47	4,83	0,00	2,09	0,06
BC	0,20	1,10	3,11	8,56	7,93	1,85	0,27	2,67	0,59	0,00	2,66	0,13
CM	0,31	0,55	0,91	4,72	6,38	5,63	5,00	7,14	0,29	0,00	2,54	2,94
CM Br	0,65	1,57	4,06	8,86	7,26	11,17	8,09	17,70	1,15	0,00	11,14	11,83
CM Di	0,46	0,50	0,61	5,68	8,75	2,83	8,13	4,26	0,97	0,00	1,90	1,15
CM Gr	0,30	0,21	0,20	6,79	4,97	5,89	3,63	3,62	0,21	0,00	2,53	5,38
CM Pr	1,44	0,28	4,81	5,37	4,49	3,77	3,33	2,49	0,14	0,00	4,69	2,47
M	0,58	3,04	0,43	0,93	1,04	0,72	0,00	0,44	8,60	0,00	5,52	0,06
TP	0,06	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,06	3,29	0,09	0,00
T	1,90	2,06	0,90	3,19	2,27	2,33	1,37	2,47	3,84	0,00	2,60	1,11
WH	0,18	0,17	0,07	11,92	10,39	3,15	6,87	4,15	0,08	0,00	3,06	2,43

Figure B.18.: Mismatch Condition (MM) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 5 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	93,66	7,88	0,75	6,21	4,65	1,72	0,44	4,05	4,83	0,00	9,25	0,43
B	1,21	90,74	0,90	1,73	1,64	0,18	0,00	0,30	6,73	0,00	3,15	0,01
BC	0,14	1,31	91,12	25,81	23,51	2,30	0,17	4,19	0,72	0,00	4,60	0,03
CM	0,19	0,42	0,92	82,38	51,60	25,65	22,70	26,33	0,12	0,00	7,16	8,11
CM Br	0,34	1,52	2,99	87,77	89,30	16,64	9,23	50,79	0,66	0,00	16,47	17,17
CM Di	0,28	0,43	0,85	70,13	49,72	87,61	19,25	20,24	1,24	0,00	5,97	1,49
CM Gr	0,25	0,11	0,07	81,64	36,90	19,60	92,33	11,65	0,11	0,00	6,29	26,70
CM Pr	3,27	0,57	5,68	76,42	57,67	13,49	8,31	85,58	0,36	0,00	30,97	8,88
M	0,33	2,83	0,19	0,43	0,49	0,31	0,00	0,19	84,32	0,00	4,36	0,01
TP	0,02	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,02	94,87	0,03	0,00
T	5,47	3,85	1,19	19,26	13,12	10,33	1,74	9,94	28,66	0,00	91,76	1,79
WH	0,07	0,05	0,02	40,51	25,17	5,16	19,54	7,29	0,02	0,00	4,85	94,00

Figure B.19.: Mismatch Condition (MM) mean results for 16 MFCC + 6 LPC (4,5 dB SNR) for 10 s. of testing data

Train / Test	AC	B	BC	CM	CM Br	CM Di	CM Gr	CM Pr	M	TP	T	WH
AC	3,13	3,46	0,75	3,71	2,69	1,59	0,56	2,22	2,07	0,00	4,92	0,46
B	1,07	5,87	0,98	1,72	1,64	0,25	0,00	0,44	4,06	0,00	2,07	0,04
BC	0,20	1,04	2,71	7,38	6,97	1,41	0,23	2,17	0,59	0,00	2,02	0,11
CM	0,24	0,48	0,81	4,42	5,95	5,39	4,44	6,73	0,25	0,00	2,34	2,62
CM Br	0,55	1,49	3,86	6,84	6,10	10,63	7,81	15,00	0,99	0,00	10,60	11,44
CM Di	0,41	0,45	0,51	4,56	8,11	2,62	8,03	3,88	0,82	0,00	1,66	0,66
CM Gr	0,22	0,21	0,17	6,47	4,47	4,09	2,94	2,55	0,21	0,00	2,29	4,12
CM Pr	1,15	0,24	4,05	4,68	3,55	3,85	3,16	1,47	0,14	0,00	3,50	2,76
M	0,58	2,83	0,43	0,86	0,92	0,62	0,00	0,44	8,61	0,00	5,11	0,04
TP	0,06	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,06	3,05	0,09	0,00
T	1,83	1,74	0,79	2,57	2,01	2,01	1,37	1,90	2,62	0,00	2,26	4,18
WH	0,15	0,14	0,05	11,52	9,51	2,31	6,50	4,01	0,08	0,00	2,85	2,19

Figure B.20.: Mismatch Condition (MM) std results for 16 MFCC + 6 LPC (4,5 dB SNR) for 10 s. of testing data

Bibliography

[1] SFB. Collaborative Research Center 588 "humanoid robots - learning and cooperating multimodal robots". [Online]. Available: http://www.sfb588.uni-karlsruhe.de/about/

[2] DFG. Deutsche Forschungsgemeinschaft / german science foundation. [Online]. Available: http://www.dfg.de/

[3] KIT. Karlsruhe Institut of Technology. [Online]. Available: http://www.kit.edu/

[4] S. B. Davis and P. Mermelstein, "Readings in speech recognition," A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, pp. 65–74.

[5] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *In International Symposium on Music Information Retrieval*, 2000.

[6] D. O'Shaughnessy, *Speech Communication: Human and Machine*. New York: IEEE Press, 2000.

[7] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[8] M. Madry-Pronobis, Master's thesis, School of Electrical Engineering, Royal Institute of Technology, May. 2009.

[9] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, May 2001.

[10] A. Carrion Isbert, *Diseño Acústico de Espacios Arquitectónicos*. Ediciones UPC, 1998.

[11] C. C. J. M. Hak, "The effect of the acoustics of sound control rooms on the perceived acoustics of a live concert hall recording," in *Proceedings of the 11th WSEAS international conference on Acoustics music theory applications*, ser. AMTA'10. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 55–60.

[12] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1973.

[13] B. Scherrer, "Gaussian mixture model classifiers," 2007.

[14] A. Swerdlow, T. Machmer, B. Kühn, and K. Kroschel, "Robust sound source identification for a humanoid robot," in *Electronic Speech Signal Processing (ESSV2008)*, Frankfurt, Germany, Sep. 2008.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Jorunal of the royal statistical society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] S. Borman, "The expectation maximization algorithm – a short tutorial," 2004.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000.

[18] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, Jan. 1970.

[19] T. C. W. Landgrebe, D. M. J. Tax, P. Paclík, and R. P. W. Duin, "The interaction between classification and reject performance for distance-based reject-option classifiers," 2005.

[20] D. M. J. Tax and R. P. W. Duin, "Growing a multi-class classifier with a reject option," *Pattern Recogn. Lett.*, vol. 29, pp. 1565–1570, July 2008.

[21] T. L, D. M. J. Tax, P. Paclík, R. P. W. Duin, and C. Andrew, "A combining strategy for ill-defined problems," 2008.

[22] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, Sep. 1998.

[23] D. M. J. Tax, Master's thesis, Advanced School for Computing and Imaging, Jun. 2001.

[24] A. J. Smola, B. Schölkopf, and K.-R. Müller, "The connection between regularization operators and support vector kernels," *Neural Netw.*, vol. 11, pp. 637–649, June 1998.

[25] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2758–2765, 1997.

[26] B. Schölkopf, J. C. Platt, J. S. Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," Microsoft Research, Tech. Rep., 1999.

[27] Beyerdinamic. Mce 60 - technical specifications sheet. [Online]. Available: http://europe.beyerdynamic.com/shop/media//datenblaetter/mce60_data_en.pdf

[28] Motu. Motu 8pre firewire audio interface. [Online]. Available: http://www.motu.com/products/motuaudio/8pre

[29] B. Maguire, Master's thesis, Rutgers University, May. 2008.

[30] R. F. National Institute of Applied Sciences (INSA), ASI Departament(Architechture des Systemes d'information). Svm and kernel methods matlab toolbox. [Online]. Available: http://asi.insa-rouen.fr/~arakotom/toolbox/index.html

[31] S. Degroeve, K. Tanghe, B. D. Baets, M. Leman, and J. pierre Martens, "A simulated annealing optimization of audio features for drum classification," in *In Proceedings of the 6th International Conference on Music Information Retrieval*, 2005, pp. 482–487.

[32] S. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman, "Speaker identification using mel frequency cepstral coefficients," in *3rd Proceedings of International Conference on Electrical and Computer Engineering*, December 2004, pp. 565–568.