# Intelligibility Based Automatic Volume Control for Public Address Systems

THESIS

Submitted in partial fulfilment of the
requirements for the degree of

MASTER OF SCIENCE

In

EMBEDDED SYSTEMS

By

Johannes Adriaan (Han) Oosterom

Student number: 1176315

Thesis Defence: 5th July 2011


Signal & Information Processing Lab
Department of Mediamatics
Faculty of EEMCS,
Delft University of Technology
Delft, the Netherlands


Bosch Security Systems
Kappitelweg 10 Breda
The Netherlands

# Preface

This thesis presents the results of my master thesis research conducted at Bosch Security Systems in Breda (8 November 2010 to 5 July 2011). The topic was designing an automatic volume control for a public address system. The research was done in cooperation with the Signal & Information Processing Lab of the TU-Delft. The result is an automatic volume control algorithm, which bases its control on the intelligibility and loudness of the public address system.

COMMITTEE MEMBERS

Dr.ir. A.J. den Dekker (3ME DCSC – TU Delft)

Dr.ir. A.J. van Genderen (EWI CE – TU Delft)

Dr.ir. R. Heusdens (EWI MSP – TU Delft)

Ir. H.S.P van der Schaar (Bosch Security Systems)

My thanks go out to my daily supervisors: Hans van der Schaar, Richard Heusdens and Richard Hendriks, for their help and their time, which made it possible for me to reach the presented result. I would also like to thank my colleagues at Bosch who gave useful feedback and input. Special thanks to the colleagues who I shared a room with: Johan van Iersel, Jan Meijer and Chen Tchang for their valuable and interesting input and output during the coffee brakes. Last, but surely not least, I would like to thank my wife and daughter for their patience and excellent support.

Han Oosterom

# Table of Contents

**Preface** ..... iii

**Table of Contents** ..... v

**List of Figures** ..... vii

**Definitions, symbols and abbreviations** ..... viii

**Abstract** ..... ix

**1. Introduction** ..... 1

1.1. PROBLEM STATEMENT ..... 2

1.2. PROJECT DESCRIPTION ..... 3

1.3. THESIS OUTLINE ..... 4

**2. Prior Art** ..... 5

2.1. EXISTING SOLUTIONS ..... 5

2.1.1. Noise extraction ..... 5

2.2. INTELLIGIBILITY ..... 8

2.3. STOI ..... 9

**3. Performed Work** ..... 12

3.1. THEORETICAL ANALYSIS ..... 12

3.1.1. Theoretical model ..... 12

3.1.2. Behaviour in the mean ..... 14

3.1.3. Optimal point of control ..... 17

3.1.4. Optimal Controller ..... 19

3.1.5. Performance Evaluation of a Specific Solution ..... 20

3.2. REQUIREMENTS ..... 22

3.2.1. Environmental requirements ..... 22

3.2.2. Algorithm requirements ..... 22

3.3. PROPOSED SOLUTION ..... 23

3.3.1. STOI adaptations ..... 24

3.3.2. SD ..... 25

3.3.3. AVC ..... 26

3.3.4. Silence mode ..... 28

3.3.5. BGM ..... 28

3.3.6. Possible extensions ..... 29

3.4. SIMULATION MODEL ..... 30

3.4.1. Live set-up simulation model ..... 30

3.5. TESTS ..... 32

3.6. SIMULATION TESTS ..... 32

3.7. LIVE TESTS ..... 33

3.8. TEST RESULTS ..... 33
3.8.1. Simulation ..... 33
3.8.2. Live Tests ..... 34

**4. Conclusions ..... 35**

4.1. SUGGESTIONS FOR FURTHER RESEARCH ..... 36

**5. References ..... 37**

**6. Appendices ..... 39**

A. MATHEMATICAL PROOFS ..... 39
A.I Derivation of derivative with respect to the Gain ..... 39

B. REQUIREMENTS ..... 40
B.I Context ..... 40
B.II AVC ..... 40
B.III Normal operation during Call ..... 41
B.IV Operation between speech ..... 42
B.V Background music operation ..... 42
B.VI Operation during Transitions ..... 42
B.VII General requirements ..... 43

C. TEST DOCUMENT ..... 44
C.I Simulation Setup: ..... 44
C.II Test List: ..... 44
C.III Live setup: ..... 47
C.IV Test proposed for future work ..... 48

# List of Figures

Figure 1-1 Locations with a PA system ..... 1
Figure 1-2 PA Model ..... 2
Figure 1-3 PA Model with AVC ..... 3
Figure 2-1 Noise extractor using a (adaptive) Filter ..... 7
Figure 2-2 STOI model [1] ..... 11
Figure 3-1 Simplified signal model ..... 13
Figure 3-2 mean d-factor (dot) as a function of the gain with mapping (red-line) ..... 16
Figure 3-3 mean intelligibility (dot) as a function of the gain with mapping (red-line) ..... 17
Figure 3-4 plot of ratio (3-14) (top) and the intelligibility Eq. (3-7) (bottom) ..... 19
Figure 3-5 Optimal Controller ..... 20
Figure 3-6 Basic control Model ..... 23
Figure 3-7 Input model ..... 23
Figure 3-8 State diagram of the controller ..... 24
Figure 3-9 Signal Detection ..... 25
Figure 3-10 Asymmetric proportional control ..... 26
Figure 3-11 Asymmetric PI-controller for AVC ..... 27
Figure 3-12 Possible extension using intelligibility based AVC ..... 29
Figure 3-13 Live set-up ..... 31
Figure 6-1 Mode of operation state diagram ..... 41
Figure 6-2 Setup for live tests ..... 47

# Definitions, symbols and abbreviations

**DEFINTITIONS:**

| | |
|---|---|
| **Audible**: | The music is heard, and not masked by noise |
| **Ambient Noise**: | Any other type of acoustical signal in the room that is not directly produced by the PA-system. |
| **Call**: | a message played by the PA system |
| **Intelligible**: | The words of a message are recognised and understood |
| **Sensing Microphone**: | Generally an Omni-directional microphone placed within the room to capture the ambient noise. |
| **Zone**: | Area within the building/space that receives/plays the same calls |

**SYMBOLS:**

In equations the following mathematical notation is used:

| | |
|---|---|
| $\|x\|$ | is the magnitude of a complex variable |
| $\mathbf{X}$ | Bold face variables are vectors |
| $\mu_{\mathbf{X}}$ | $\mu$ is used as a sample mean of the corresponding vector |
| $X^*$ | The uppercase star is used as the complex conjugate of a complex variable |
| $X$ | Capital letters denote frequency domain magnitudes |
| $\mathbf{X}^{\mathrm{T}}$ | is the transpose of the vector **X** |
| $\|\mathbf{X}\|$ | is the l2-norm of a vector |
| $d$ | is the intelligibility correlation factor calculated by Eq. (2-6) |
| $d_r$ | is the intelligibility calculated using Eq. (2-7) |
| $G$ | gain applied to the clean speech |

**ABBREVIATIONS:**

| | |
|---|---|
| **AVC** | Automatic Volume Control |
| **BGM** | Background music. This is music played by the public address system to support a certain atmosphere |
| **DFT** | Discrete Fourier Transform |
| **DSP** | Digital Signal Processor |
| **FPGA** | Field Programmable Gate Array |
| **Fs** | Sampling Frequency |
| **PA** | Public Address system |
| **PID-controller** | Proportional-Integral-Derivative controller |
| **SD** | Signal Detection |
| **SNR** | Signal to Noise Ratio |
| **STOI** | Short-Time Objective Intelligibility |

# Abstract

*To convey messages to the public, public address systems (PA) are installed in buildings and at venues. These messages generally contain important information for the listener. This information has to come across well, i.e. the message should be intelligible. Because the environment, and mainly the background noise, can change over time, it is important for a public address system to adapt accordingly, so that the intelligibility of the messages is maintained. To maintain the intelligibility automatic volume control algorithms are used. In current solutions these algorithms adapt the volume to maintain the signal to noise ratio at a constant level. Such approaches require acquiring information about the noise from a sensing microphone. The difficulty in this is that the sensing microphone not only captures the noise, but also the signal coming from the PA itself, including its echoes and reverberations.*

*To avoid the signal separation problem, the proposed solution directly analyses the intelligibility of the message using the signal from the sensing microphone. For this an objective intelligibility method was used, that analyses correlations between the original clean message and the distorted message, from the microphone. Using the found intelligibility, the volume is controlled to maintain intelligibility. However, because maximum intelligibility occurs at the maximum volume of the PA system, before the signal starts deforming, maintaining intelligibility alone is not enough. Loud PA systems are perceived to be annoying especially if the background noise is low. That is why the proposed solution limits the loudness of the PA system in combination with maintaining the intelligibility.*

# 1. Introduction

Figure 1-1 Locations with a PA system

The purpose of a Public Address system (PA) is to convey messages to the listeners in a certain area. These public areas can range from a train station, to a cruise liner or an office-building. Figure 1-1 illustrates some more examples of buildings or locations that house a PA system. A message, from such a system, can be a platform/gate change, the building closing, evacuation instructions, etc. A PA system message, carrying information, is from now on referred to as a call. For calls it is of primary importance that the message comes across and is well understood, i.e. is intelligible.

A simple scheme of a PA is shown in Figure 1-2. In this model the following components can be distinguished: A call source, an amplifier and a loudspeaker. The call source can generate two different signals, the aforementioned call and background music (BGM). Where the call conveys a specific message, the BGM is used to set or support a certain atmosphere.

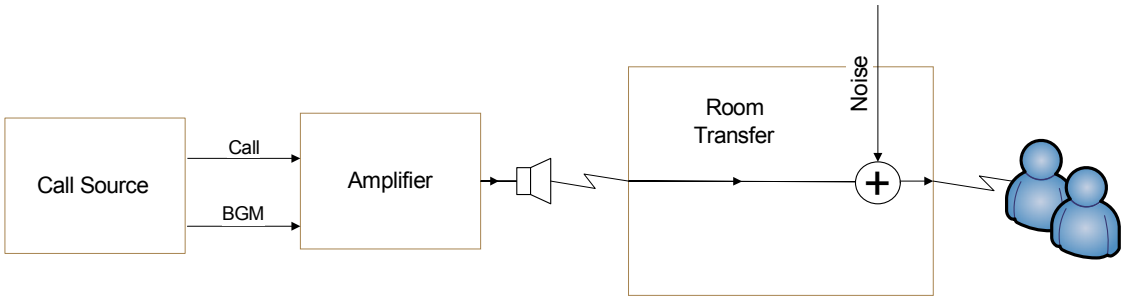

Figure 1-2 PA Model

The room transfer models the room characteristics, i.e., the acoustic path from the loudspeaker to the point of interest in the room. In the room, echoes, reverberations and noise distort the original clean signal. Both the room characteristics, modelled by a room impulse response, and the noise are unknown beforehand.

## 1.1. PROBLEM STATEMENT

For a PA system it is not only important that the messages are intelligible, it should also be pleasant to listen to. For a PA system to be pleasant to listen to, it should not be too loud. A loud PA system can startle the listener when a call commences, or for example disrupt the night rest of people living close to the building, containing the PA system. The perception of loudness is a function of the level of the background noise. If the noise level is high a higher volume is tolerated compared to quiet moments ([8]) the intelligibility however decreases if more noise is present ([17]). Both the perception of loudness and the intelligibility decrease with an increasing background noise level.

In order to maintain the intelligibility and limit the loudness of a PA system a volume control is necessary. A volume control should keep the volume at such a level that the messages are intelligible but not too loud. Because the noise level varies over time, it is desirable that the volume control is an Automatic Volume Control (AVC) that can automatically adjust depending on acoustic changes in the environment.

Figure 1-3 illustrates where the AVC fits into the PA model (Figure 1-2). The AVC makes use of at least one sensing microphone to sense the signal in the room. Moreover the AVC module receives the clean call or background music signal. Given these signals, the AVC, should calculate the gain that the system should operate at. On top of the problem of maintaining the intelligibility and limiting the loudness a new problem is introduced by introducing a loop in the system. With this closed loop there is a risk that the system might become unstable. The sensing microphone does not only record the noise level, it also records the original signal, played through the loudspeaker, together with its echoes and reverberations added in the room. The risk is in the fact that if the volume control is not able to distinguish sufficiently well between the noise and its own signal that the volume could be set at a too high level, or the system could go out of control by continuously increasing the gain because it interprets the increasing original signal as increasing background noise. In this thesis this problem is referred to as the feedback problem.

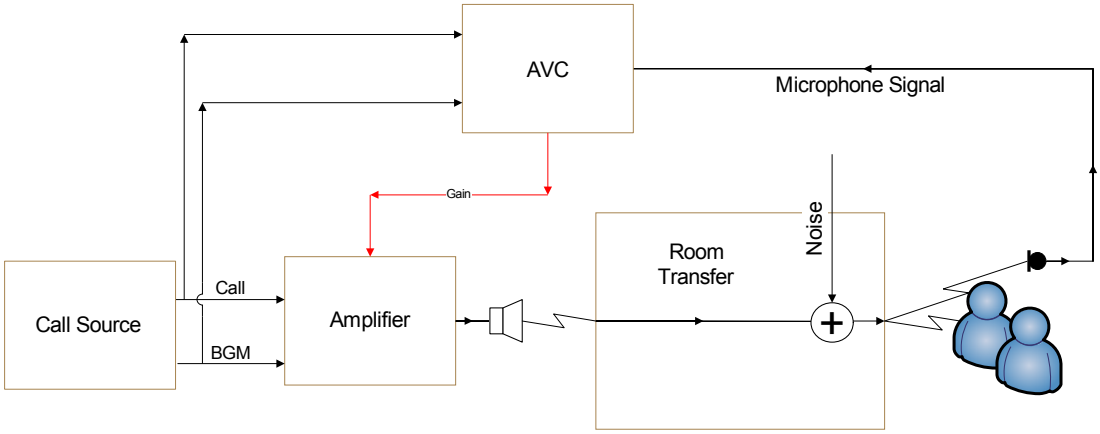

Figure 1-3 PA Model with AVC

Some existing solutions only adapt the volume in-between calls (for example [10] and [13]), i.e., if the noise level changes during a call, the playback level is not modified according to intelligibility and comfort. Hence, this can lead to a playback level that is either too loud or too low. Once a call is started, the gain is frozen and starts to re-adapt after the echoes and reverberations have died out. This approach avoids the feedback problem by only sensing at moments when the microphone captures only noise. However, this will also slow down the speed at which a PA system can react on a changing acoustical environment. Furthermore, the way in which AVCs usually adjust the playback level is often based on heuristic grounds like assuming that feedback is always the same percentage of the energy or assuming that changing the volume very slowly will avoid the feedback problem([11], [12]). Other solutions rely on measures like the signal-to-noise ratio (SNR) (i.e. [14], [17]) to determine an appropriate gain. But this requires finding the noise only signal/energy.

## 1.2. PROJECT DESCRIPTION

The goal of this master project is to design an AVC that can continuously adapt, even during a message. Moreover the system has to have a low complexity and be implemented on a FPGA or DSP.

In this master project I propose to make use of an objective intelligibility model in order to automatically control the volume. The proposed system consists of an algorithm to calculate the intelligibility of the call, and a volume control that increases the gain if the intelligibility is too low and decreases the gain when intelligibility is satisfactory but the PA system is too loud. By using a correlation based intelligibility measure instead of the SNR, the sensing microphone signal does not have to be filtered to separate the noise from the total signal. Moreover there is a direct link to the most important requirement of a PA system, to 'maintaining the intelligibility'. Using an objective intelligibility measure also adds new functionalities to existing PA systems like automatic call repetition when a call was unintelligible.

Special attention is paid to having a fast reaction speed to changes in noise level. A good example of a fast varying noise source is the train arriving at the train station.

## 1.3. THESIS OUTLINE

This thesis has the following structure: Chapter 2 starts by explaining which solutions already exist as automatic volume controllers and how they work. This chapter also explains some related research on e.g., objective intelligibility measurements methods, which is of importance for the designed system. Then Chapter 3 explains how the new volume control is designed, starting off with a theoretical analysis (Section 3.1) where the effects of the chosen control method are analysed and justified. This section also shows how it can be verified that the new algorithm is performing better than other options. Followed by the requirements (Section 3.2) set before designing the controller. Section 3.3.3 describes how the controller works and Section 3.4 describes how and why a simulation model was build. This chapter closes of with the tests (Section 3.5), with the results, that where performed to verify the functioning of the system and whether the system requirements where met.

Chapter 4 successively closes off with the evaluation and conclusions, followed by suggestions for future research (Section 4.1).

# 2. Prior Art

Because AVC is not a novel feature, it is important to initially analyse the existing solutions before presenting the proposed solution. In this chapter a brief overview of existing solutions for AVC is given, as well as some necessary background information. In Section 2.1 a number of important solutions are described, analysed and compared. Section 2.2 and 2.3 gives a short description on what is understood about intelligibility and how it can be measured. The latter section discusses the specific model that was the basis for the proposed solution.

## 2.1. EXISTING SOLUTIONS

In general, all AVCs designed to date, have a similar underlying idea. However, the result is realized in different manners. The goal of all the AVCs is to regulate the volume on the basis of the level of the background noise. The underlying thought is that the louder the background noise, the higher the volume should be. This can be characterized by a system in which the SNR is kept at a constant level (i.e. [14], [17]).

The main problem with this approach is that the noise is not available as a separate signal. In general, a sensing microphone is used that is placed within the space where the AVC should work. The sensing microphone however captures not only the noise but also the call played by the PA system, with its echoes and reverberations. Different implementations use varying methods to separate the noise from the reverberant signal, the most commonly used methods will be discussed in the following section. If the separation is not successful, there is a risk that the system could go into a gain chase where the system keeps increasing its gain because it interprets its own signal as noise as well. Generally there is an upper limit defined for the gain, so such a system would set its maximal gain and get stuck there. A maximal gain is defined to protect the system and the ears of the listeners. This gain chase is the aforementioned feedback problem.

### 2.1.1. Noise extraction

The most basic method to sense the noise, is to measure only at moments when there is no call or remainder thereof, in the form of echoes or reverberations ([10], [13]). Clearly, the advantage of this approach is that the AVC can be kept relatively simple and does not have to deal with the problem that in the ambient signal, noise is mixed with the call and its reverberations. However, this is at the high cost of not being able to adapt the volume during a call. For maintaining intelligibility it is of great importance to adapt during the call. I again refer to the example of the train coming in. If the train comes in right after the start of the call, the message is lost. A similar thing happens when the train leaves during a call. Then the PA system plays the call very loudly while the background noise is relatively quiet. These types of systems only work well for very slowly changing background noise that does not change significantly during a call. This approach to extract the noise also has the problem that it is highly dependent on the acoustics of the room. If the echoes, and reverberation time, are short the system can adapt more frequently than a system working in a large room with a lot of echoes. Generally this time

is stored in a parameter with a certain safety margin ([10]). This implies that if the echo and reverberation time reduce, the system would still react at the same speed while a higher speed would be possible. The reduction could be caused by a higher damping in the room. This could be a result of more people being present.

The second group of solutions focuses on acquiring an indication of the noise energy. In [12] this is done by taking the energy of the sensed signal ($E_s$) and subtracting the energy of the amplified call signal ($E_{Gx}$).These energies are calculated over a specific time frame to track changes over time. The SNR could then be calculated in the following manner:

$$SNR = \frac{E_{Gx}}{E_s - E_{Gx}}, \tag{2-1}$$

Such an approach completely disregards the echoes, and when the system would be used, in a room where these echoes are strong, the feedback problem would still occur. In [19], [20] an extra component is added to the SNR calculated in Eq. (2-1) to compensate for these echoes. However, this rather heuristic method assumes that this is a constant value or percentage of the original amplified call. This solution can also not cope with changes in the room acoustics like the previously described solution. The SNR calculated in such a system can be calculated in the following manner where $E_{er}$ is an estimation of the energy of the echoes and reverberations:

$$SNR = \frac{E_{Gx}}{E_s - E_{Gx} - E_{er}}, \tag{2-2}$$

Within both these solutions it is important that the sensed signal is time aligned with the amplified signal so that the energy subtraction occurs over the right fragment. In case there is a misalignment the SNR is again not calculated correctly because too little or too much of the sensed signal is subtracted. The advantage this solution is that it is very simple.

Other solutions perform some kind of filtering operation in order to filter out the original signal from the sensed mixed signal to estimate the noise. Within these solutions with a filter, there are also different levels of accuracy. In for example [21] the assumption is made that the noise energy is mainly present in the low frequency range where there is hardly any speech present. Such an approach overlooks noise sources in other frequency ranges that could seriously degrade the intelligibility of the call. Other solutions make the assumption that the noise is white, and thus knowing the noise level in the lower frequencies is enough to determine the level for the higher frequencies.

The more robust approach uses a (adaptive) filter, which removes the original signal with the most important echoes from the sensed signal. These approaches are based on the principle of an echo or feedback canceller ([4]). The filter tries to model the room impulse response, with as goal to be able to recreate the exact call components present in the room and subsequently subtract them from the sensed signal.

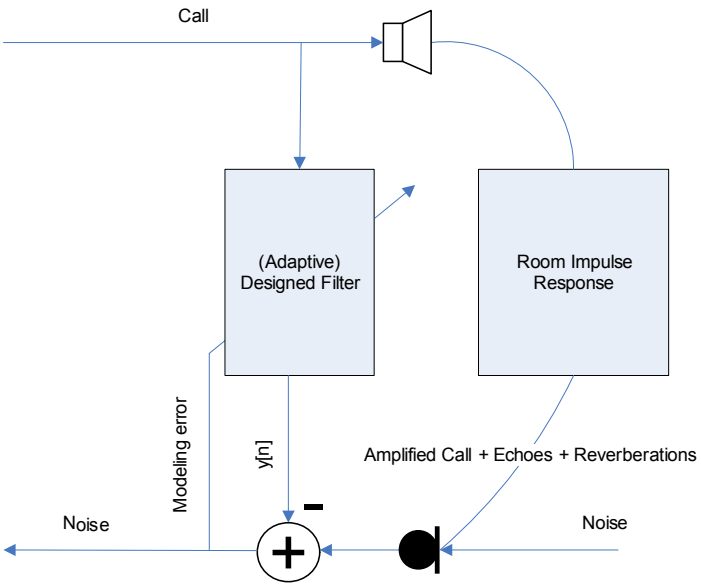


Figure 2-1 Noise extractor using a (adaptive) Filter

The model shown in Figure 2-1 gives an illustration on how such a system works. It also shows the problem with such a system. That is, the modelling error is contaminated by the noise. If the noise is white and uncorrelated with the call this is generally not a problem. But if the noise is coloured the converging behaviour of the adaptive filter is influenced ([4]). In [2] a more extensive analysis is done in the performance and complexity of adaptive filter algorithms that can be used for such a system.
The main conclusions in the conducted literature study are:

- To model the impulse response of a large room accurately a very high order filter is needed. The length can be limited by making an estimation of the energy in the echo tail of the call([22])
- Algorithms that adapt quickly generally have a high complexity (quickly adapt the filter to changes in room acoustics). For example the RLS algorithm converges faster than the LMS algorithm but it is also an order of magnitude more complex.
- There are number of trade-offs in designing an adaptive filter that make it difficult, to impossible, to find a filter that performs well in every aspect. Low complexity, high speed, accurate impulse response modelling and quick adaptations to changes in the impulse response are opposing requirements.

Automatic volume controls do exist that use this approach of adaptive filtering. However, they generally have a relatively slow adaptation rate, with a limited gain step every few seconds. Moreover these algorithms are sensitive to the environment in which they are installed. In many cases, as mentioned before, they are also, in some sense, sensitive to the characteristics (colour) of the signals used (call and noise)([4]). These factors all influence the robustness of the algorithm and the system, and limit the implementability. Moreover it could be questioned whether extracting the noise signal itself is not too excessive, to successively only calculate the SNR in order to steer the playback level of the PA system.

Moreover, an extra requirement is generally made that the microphone is placed at a location in the room where the call signal component is as small as possible.

## 2.2. INTELLIGIBILITY

As already mentioned in the introduction the goal of an AVC is to improve the intelligibility. Therefore, it would be of interest to somehow quantify intelligibility, because this would be a direct feedback on how well the system performs. Intelligibility is generally, and most accurately, quantified through conducting listening tests. In such a listening test a set of listeners gets to listen to distorted (noisy) speech samples. The intelligibility score is then the percentage of words recognised correctly.

The problem with these so called subjective tests is that they are time consuming and costly. Using such a way of measuring in combination with an AVC would imply that an expert listener would be deployed to listen to the calls, and when they become unintelligible, to increase the volume. It is of course clear that this is an infeasible, but probably accurate solution to the intelligibility part of AVC challenge.

Because of the cost and time involved in conducting subjective tests, methods have been developed to mimic these subjective intelligibility tests, so-called objective intelligibility methods. Such objective methods use an algorithm to analyse the intelligibility of a certain speech fragment. A number of objective algorithms are compared in [5]. From this comparison it becomes clear that not all objective models have a high correlation with the subjective test results. This is something that has to be taken into account when using one of these models. Moreover some of these algorithms require a great number of parameters to be determined ([1]). The general use of such algorithms is to analyse the performance of noise suppression or speech coding algorithms. This thesis analyses the possibility of using one of these algorithms inside an AVC for PA systems.

## 2.3. STOI

An algorithm that has been shown to be a good predictor of subjective intelligibility is the Short-Time Objective Intelligibility (STOI) ([1]). This is an algorithm that computes the correlation between 30 Discrete Fourier Transform (DFT) magnitudes in a certain one-third octave band and over an approximately 400ms long time frame. The clean speech and the noisy speech are compared by means of correlation in order to analyse how intelligible the degraded (noisy) speech is. This turns out to be an accurate predictor of the intelligibility for a certain speech fragment. Figure 2-2 shows the basic signal flow diagram of the STOI algorithm.

The STOI algorithm works with two time aligned signals namely the clean speech signal $x$ and the degraded speech signal $y$. These signals are resampled at 10 kHz and cut into 256 sample Hann-windowed time frames, with a 50% overlap. The time frames in $x$ and $y$ that do not contain speech energy are discarded. The individual time frames are zero padded to become 512 samples long before applying a DFT.

Then, a one-third octave band analysis is performed, grouping the DFT-bins into 15 one-third octave bands. Let $\hat{x}(k,m)$ denote the $k^{th}$ DFT-bin of the $m^{th}$ frame of the clean speech. The norm of the $j^{th}$ one-third octave band, referred to as a Time Frequency-unit (TF-unit), is defined as,

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} \left|\hat{x}(k,m)\right|^2}\,, \tag{2-3}$$

where $k_1$ and $k_2$ denote the one-third octave band edges. The TF-units for the degraded speech are obtained similarly, and are denoted by $Y_j(m)$.

As mentioned before, STOI compares the temporal envelopes of the clean and degraded speech in short-time regions by means of a correlation coefficient. The following vector can be defined to denote the short-time temporal envelope of the clean speech,

$$\mathbf{X}_{j,m} = \begin{bmatrix} X_j(m-L+1) \\ X_j(m-L+2) \\ \vdots \\ X_j(m) \end{bmatrix}, \tag{2-4}$$

where $L$ denotes the number of consecutive TF-units that are grouped into one vector. For the STOI algorithm $L$=30, which equals an analysis length of 384ms (at a sampling frequency of 10kHz). For the degraded speech, the vector $\mathbf{Y}_{j,m}$ is obtained in a similar manner. Before comparison $\mathbf{Y}_{j,m}$ is first normalized and clipped. For a more detailed description and motivation of the normalization and clipping procedure, also shown in Figure 2-2, see [1]. The intermediate intelligibility, $d_{j,m}$, is then defined as the correlation between the two vectors,

$$d_{j,m} = \frac{\left(\mathbf{X}_{j,m} - \mu_{\mathbf{X}_{j,m}}\right)^{\mathrm{T}} \left(\mathbf{Y}_{j,m} - \mu_{\mathbf{Y}_{j,m}}\right)}{\left\|\mathbf{X}_{j,m} - \mu_{\mathbf{X}_{j,m}}\right\| \left\|\mathbf{Y}_{j,m} - \mu_{\mathbf{Y}_{j,m}}\right\|}, \tag{2-5}$$

where $\mu_X$ and $\underline{\mu_Y}$ are the sample mean of the vector $\mathbf{X}$ and $\mathbf{Y}$ respectively.
Finally the average of the intermediate intelligibility over all bands and al frames is calculated,

$$d = \frac{1}{JM} \sum_{j,m} d_{j,m}\,, \tag{2-6}$$

Where $M$ represents the total number of frames and $J$ the number of one-third octave bands. The STOI algorithm calculates a intelligibility factor $d$, this factor can be mapped onto a realistic (subjective) intelligibility percentage using the following mapping.

$$d_r = \frac{100}{1 + \mathrm{e}^{-13.1903d + 6.5192}}\,, \tag{2-7}$$

To make this algorithm suitable for use in an AVC for a PA system, a number of small modifications are made. These are discussed in Section 3.3.1.

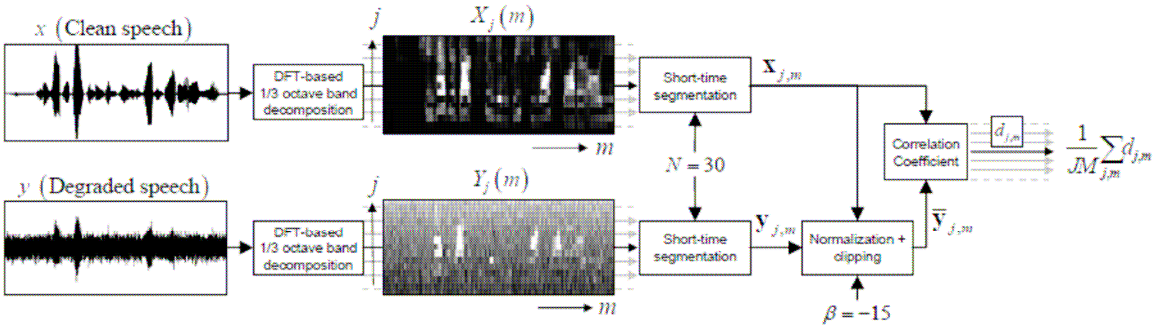


Figure 2-2 STOI model [1]

This algorithm is chosen for use in the control algorithm because it has the following advantages over a number of other algorithms:

- Low algorithmic complexity
- A lot of parallelism (useful for hardware implementation)
- High Correlation with subjective tests ([1] and [5])
- No extensive parameter set-up [1]
- Can be used in real-time because of frame-based analysis
- Relatively short analysis window
- Has the possibility of working for music (audibility)
- A correlation factor is calculated per one-third octave band this opens the door to control the gain per one-third octave band as well.

# 3. Performed Work

This chapter describes the steps that are taken in designing, and verification a novel solution for AVC. Firstly, a theoretical analysis of the problem and the intelligibility model used is given. Then a description is given on how the actual AVC works in achieving the predefined goals. The chapter ends with the presentation of a simulation model and the tests results achieved using this model.

## 3.1. THEORETICAL ANALYSIS

With any design it is of importance to analyse whether the theoretical basis is firm, and analyse where difficulties may arise. It was discussed in the previous chapter that an interesting way of measuring the performance of the PA system is by using the short-time objective intelligibility measure, STOI.

In this section an analysis is performed on how this intelligibility model relates to the given control parameter (the gain, in a PA system). To be able to design a system that measures the intelligibility, and uses that to calculate an appropriate gain, this relationship has to be clear. Preferably the intelligibility should be a monotonically rising function of the gain. This would imply that if the gain is increased that the intelligibility also increases. This relationship is studied in Section 3.1.2. To be able to study this relationship, an abstract model of the STOI-model was made. Section 3.1.1 describes this theoretical model and what changes have been made with respect to the original model presented in Section 2.3. To combine both the intelligibility and the loudness into one model it is important to define an error and find an optimal point. Section 3.1.3 defines a ratio that helps in finding an interesting intelligibility level that can be used as a set-point for the controller. Using this set-point as a target, Section 3.1.4 discloses a basic controller that would be able to minimize the error. Finally, a more detailed description of the error, and how it can be used to represent both the lack of intelligibility or the system being too loud are given in Section 3.1.5.

### 3.1.1. Theoretical model

For the theoretical analysis we propose a number of simplifications to the STOI model discussed in Section 2.3. The analysis is performed on only one band of the DFT, which means that the one-third octave band separation is not made. Secondly, the non-linear effect of the clipping is left out of consideration for simplicity. When clipping is not used the normalization step is not necessary, which is therefore also left out. To simplify the equations, we introduce two new vectors that are defined as follows:

$$\mathbf{C} = \mathbf{X}_{j,m} - \mu_{\mathbf{X}_{j,m}},$$

and

$$\mathbf{Z} = \mathbf{Y}_{j,m} - \mu_{\mathbf{Y}_{j,m}},$$

where $\mu_{\mathbf{X}_{j,m}}$ and $\mu_{\mathbf{Y}_{j,m}}$ are the sample mean of the clean($\mathbf{X}$) and distorted speech($\mathbf{Y}$) vector respectively.

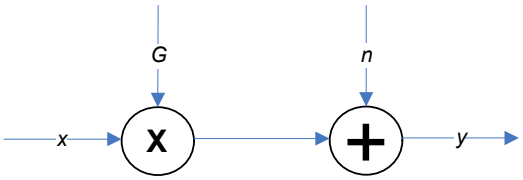


Figure 3-1 Simplified signal model

The intermediate intelligibility measure, *d* Eq. (2-5) can be rewritten in a more compact form as Eq. (3-1).

$$d = \frac{\mathbf{C}^{\mathrm{T}}\mathbf{Z}}{\|\mathbf{C}\|\,\|\mathbf{Z}\|}, \tag{3-1}$$

This equation expresses the correlation between the clean speech vector **C** and the sensed signal vector **Z**. For this analysis, one DFT-bin *k* is chosen and *m* is set equal to *L* to analyse only one frame. So Eq. (2-6) becomes unnecessary, i.e. the intermediate intelligibility is the used intelligibility measure.

The next step in the theoretical analysis is to define the distorted signal *y*. For this analysis the simplified model that is shown in Figure 3-1 is used for the sensing microphone signal *y*. Figure 3-1 is a simplified version of the room model shown in Figure 1-2. No room transfer function is taken into account, which means there are no echoes or reverberations. The noise *n* is additive.

For the microphone signal we can then use the following expression for the DFT-coefficients:

$$\hat{y} = G\hat{x} + \hat{n},$$

where *G*, is the gain applied to the current analysis frame and the '^' indicates a DFT coefficient. Because we are interested in the magnitude of this microphone signal we get the following expression for these magnitudes:

$$Y(G,i) = \sqrt{G^2\left|\hat{x}(k,i)\right|^2 + G\hat{x}(k,i)\hat{n}^*(k,i) + G\hat{x}^*(k,i)\hat{n}(k,i) + \left|\hat{n}(k,i)\right|^2}, \tag{3-2}$$

where *i* is the vector index ($i \in \{1,2,\cdots,L\}$). In Eq. (3-2), $\hat{n}$ and $|\hat{n}|$, denote the noise DFT coefficient and its magnitude in an identical manner to $\hat{x}$ and $|\hat{x}|$. The index *k* for the frequency bin is again fixed. The *-operation is used as the complex conjugate. Eq. (3-1) analysed as function of the gain is:

$$d(G) = \frac{\mathbf{C}^{\mathrm{T}}\mathbf{Z}(G)}{\|\mathbf{C}\|\,\|\mathbf{Z}(G)\|}, \tag{3-3}$$

To verify that Eq. (3-3) is monotonically rising, it should be verified that the derivative with respect to *G*, is always positive. The derivative of Eq. (3-3) is found to be:

$$\frac{\partial d(G)}{\partial G} = \frac{\left(\mathbf{C}^{\mathrm{T}}\frac{\partial \mathbf{Z}}{\partial G}\right)\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right) - \left(\mathbf{C}^{\mathrm{T}}\mathbf{Z}\right)\left(\mathbf{Z}^{\mathrm{T}}\frac{\partial(\mathbf{Z})}{\partial G}\right)}{\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}}\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}, \tag{3-4}$$

Where the derivative of **Z** is given by;

$$\frac{\partial \mathbf{Z}}{\partial G} = \frac{\partial \mathbf{Y}(G)}{\partial G} - \frac{1}{L}\sum_{i} \frac{\partial Y(G,i)}{\partial G},$$

and the derivative of $Y$ is given by;

$$\frac{\partial Y(G,i)}{\partial G} = \frac{GX^2(i) + \hat{x}(k,i)\hat{n}^*(k,i) + \hat{x}^*(k,i)\hat{n}(k,i)}{2Y(G,i)},$$

The full derivation of this derivative can be found in appendix A.I.

The correlation (intelligibility) calculated using Eq. (3-3) is not only dependent on the applied gain. In addition it is also highly dependent on the specific realizations of the clean speech and the noise. So an analytical evaluation, of whether this derivative is always positive results in the conclusion that the derivative is not always positive due to the dependency on realizations of the speech and noise process. The next step in the analysis is to analyse the mean behaviour in order to verify whether in expectation the derivative is positive. This analysis is done in the next section.

### 3.1.2. Behaviour in the mean

As mentioned before numerous noise and signal realizations can be found in which Eq. (3-3) is not monotonically rising as a function of the gain. Therefore the next step is to find out whether in expectation, Eq. (3-3) is monotonically rising as a function of the gain. Because finding a closed form expression for;

$$\mathrm{E}[d(G)] \text{ or } \mathrm{E}\left[\frac{\partial d(G)}{\partial G}\right],$$

is not straightforward, a sample mean is calculated from a large set of realizations of this process. The sample mean of the intelligibility, $\tilde{d}(G)$, can be calculated using the following expression,

$$\tilde{d}(G) = \frac{1}{K}\sum_{k=1}^{K} d(G,k) = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathbf{C}^{\mathrm{T}}\mathbf{Z}(G,k)}{\|\mathbf{C}\|\|\mathbf{Z}(G,k)\|}, \tag{3-5}$$

where $K$ is the number of iterations. The Law of Large Numbers ([3]), states that if $K$ becomes very large, that the sample mean $\tilde{d}(G)$ approaches the expected value E[$d(G)$]. In a similar manner the sample variance can also be expressed and shown to converge to the actual variance. The sample variance can be expressed by the following function:

$$\tilde{\mathrm{V}}\mathrm{ar}(d(G)) = \frac{1}{K-1}\sum_{k=1}^{K} (d(G,k) - \tilde{d}(G))^2, \tag{3-6}$$

Realizations can be created by taking a fixed clean signal $x$, and generating a new noise signal $n$ for every iteration, $k$ and for every new gain $G$. $G$ is chosen to range from some very small initial gain to maximally 1 i.e. $0 < G \leq 1$. This maximum gain is defined to limit the size of the experiments. Moreover the signal $x$ and $n$ can then be chosen in such a manner that if the maximum gain, i.e. 1, is applied, the intelligibility is also approximately 1 and the message is fully intelligible. Both $x$ and $n$ have a length of 3840 samples. 3840 samples is the result of taking, $L$=30 frames of 256 samples with a 50% overlap. The gain is applied over the complete clean signal $x$. So for every step in the process, the following two signals are generated and used:

$$x,$$
$$y = Gx + n,$$

As mentioned before, the clean signal $x$ is only generated once. The clean signal consists of a sinusoid that was amplitude modulated over time. The amplitude modulation was done because the STOI algorithm looks for correlations between, temporal changes in the spectral content of the clean signal, and distorted signal. The fundamental frequency of this sinusoid is chosen equal to the centre of the DFT-bin used for this intelligibility analysis. To ensure that the noise was also present in the analysed frequency bin, the noise was chosen to be Gaussian distributed white noise, which has on average a flat spectrum. Moreover the amplitude of the noise and the clean signal are matched in such a way that at the maximal gain the intelligibility, in expectation, is approximately 100%. The results of this analysis can be seen in Figure 3-2 (dotted-line), where the intelligibility factor $d$ is shown as a function of the gain. It is clear from this graph that in expectation $d$ is a monotonic increasing function of $G$. Figure 3-3 (dotted-line) uses the mapping given in Eq. (2-7) to express the realistic-intelligibility, as a function of the gain as well.

In [5] it is mentioned that "for additive white noise, it is widely accepted to model the psychometric-curve using a logistic curve as function of the SNR." The SNR in turn is related logarithmically to $G$. Therefore a logistic mapping would also be in place in the intelligibility-gain function. Analysing the shape of the intelligibility as function of the gain, and using the aforementioned relationship, it was observed that $\mathrm{E}[d_r(G)]$ could be mapped onto the following logistic function;

$$\mathrm{E}\left[d_r(G)\right] \approx \hat{d}_r(G) = \frac{100}{1+\mathrm{e}^{a\ln(G)+b}}, \tag{3-7}$$

This mapping is of interest to be able to perform arithmetic calculations to find the optimal control point for the AVC in the following section. The function shown in Figure 3-2 can also be mapped onto a similar function. This mapping could be applied to $\mathrm{E}[d(G)]$

$$\mathrm{E}\left[d(G)\right] \approx \hat{d}(G) = \frac{1}{1+\mathrm{e}^{a\ln(G)+b}}, \tag{3-8}$$

If this mapping is performed under the constraint that the signal should be intelligible, i.e. approximately 100% intelligibility, with $G$=1, and unintelligible with a gain close to zero, we get the following constants (Table 1):

| Function | *a* | *b* |
|---|---|---|
| (3-7) | -6.172 | -5.798 |
| (3-8) | -1.952 | -1.806 |

Table 1 Fitting Constants

The constant $b$ can be found directly by using the de maximal intelligibility when the gain is set to 1.

$$\frac{1}{1+\mathrm{e}^{b}} = d_{\max}, \tag{3-9}$$

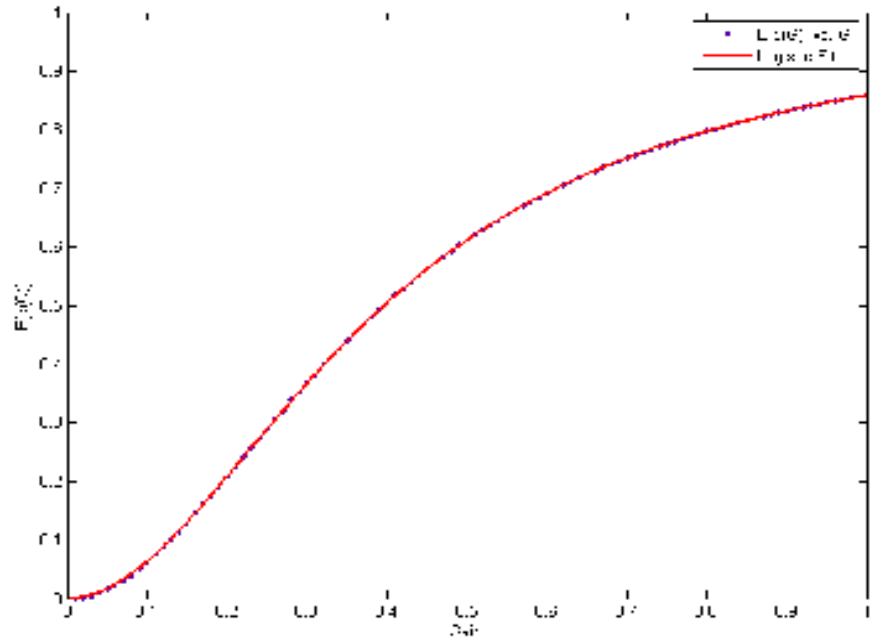

Figure 3-2 mean d-factor (dot) as a function of the gain with mapping (red-line)

$b$ can then be expressed by the following function:

$$b = \ln\left(\frac{1}{d_{\max}} - 1\right), \tag{3-10}$$

$a$ is related to the gradient at the maximal gain ($G$ =1). The derivative of Eq. (3-8) is given by;

$$\frac{\partial \hat{d}(G)}{\partial G} = \frac{-\frac{a}{G}\mathrm{e}^{a\ln(G)+b}}{\left(1+\mathrm{e}^{a\ln(G)+b}\right)^2}, \tag{3-11}$$

setting $G$ to 1, i.e., the maximum gain,

$$\frac{\partial \hat{d}(1)}{\partial G} = \frac{-a\,\mathrm{e}^{b}}{\left(1+\mathrm{e}^{b}\right)^2} \tag{3-12}$$

an expression for $a$ is obtained, that is,

$$a = \frac{-\frac{\partial \hat{d}(1)}{\partial G}\left(1+\mathrm{e}^{b}\right)^2}{\mathrm{e}^{b}}, \tag{3-13}$$

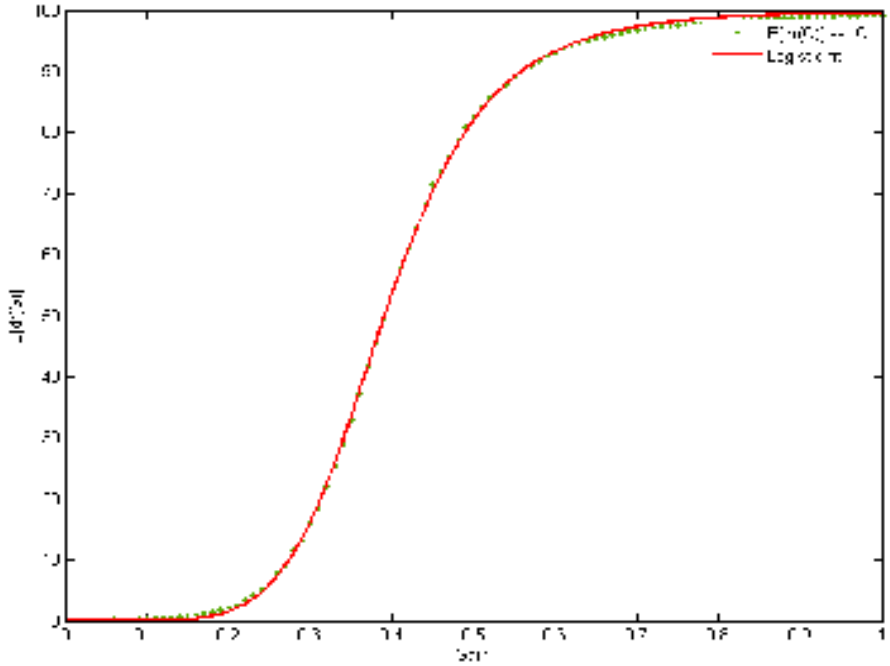

Figure 3-3 mean intelligibility (dot) as a function of the gain with mapping (red-line)

From these graphs and expressions it can be seen that the maximum intelligibility is achieved at a maximal gain. Because the problem at hand was twofold, ensuring intelligibility on the one hand and limiting the loudness on the other, it becomes important to find a point to maintain the intelligibility at, which is not at the maximum gain. At the maximum gain only the first issue, of the intelligibility, is addressed. The next section will go deeper into finding this optimal point which combines both aspects into one intelligibility level.

### 3.1.3. Optimal point of control

In the performance of the algorithm there are two factors of importance. The first is the intelligibility, the second loudness/sound power. Intelligibility is the lower-bound, which determines a lower bound for the gain to maintain a satisfactory intelligibility. Loudness is the upper bound. Once a satisfactory intelligibility is achieved it is of importance that the message is not louder than necessary i.e. the gain is not higher than strictly necessary. The behaviour of the latter is harder to quantify because this is more installation dependent. Moreover the perception of loudness is subjective and also a function of the ambient noise level ([8]). If the noise level is high, a higher PA-volume is tolerated. However, in relative silence a high volume is perceived to be annoying ([9]).

Using the monotonic relationship discovered in the previous section it is clear that if a certain level of intelligibility is chosen, adapting the gain such that the intelligibility is kept at that level is directly the minimal gain possible. What can be observed from the graph shown in Figure 3-2 and Figure 3-3 is that a certain point the gain needs to be increased more than the resulting intelligibility increases. This turning point can be found by finding the maximum of the second derivative of Eq. (3-7) or Eq. (3-8). To simplify

finding the turning point, a term, $1/G$, can be introduced. Such a term, in a sense, 'punishes' using too much gain. Using this term in Eq. (3-7) or Eq. (3-8) results in the following two ratios:

$$\frac{d(G)}{G} \quad or \quad \frac{d_r(G)}{G}, \tag{3-14}$$

These ratios can be used to approximate the location of the aforementioned turning point. $d(G)$ and $d_r(G)$ are increasing, and $1/G$ decreasing as a function of the gain. If $d(G)/d_r(G)$ increases slower than $1/G$ decreases, you are in the neighbourhood of the turning point.

The goal is to find a gain that maximizes this ratio and then find the corresponding mean intelligibility for that gain. Using the logistic model expressed in Eq. (3-8) the ratio can be expressed as the following function:

$$\frac{\hat{d}_r(G)}{G} = \frac{100}{G + G\,\mathrm{e}^{a\ln(G)+b}}, \tag{3-15}$$

To find the maximum of this function the following derivative of Eq. (3-15) has to be set to zero. In Figure 3-4 both the original function Eq. (3-7) and the ratio Eq. (3-14) are plotted. The top line is the ratio, and the bottom line the corresponding intelligibility. The axis show where the maximum occurs in the ratio, and what the corresponding intelligibility is for that gain. The derivative of Eq. (3-15) can be given by:

$$\frac{\partial\left(\hat{d}_r(G)/G\right)}{\partial G} = \frac{-100\left(1 + \mathrm{e}^{a\ln(G)+b} + a\,\mathrm{e}^{a\ln(G)+b}\right)}{\left(G + G\,\mathrm{e}^{a\ln(G)+b}\right)^2}, \tag{3-16}$$

Finding the zero of Eq. (3-16) is equivalent to finding the zero of the numerator:

$$-100\left(1 + (1+a)\,\mathrm{e}^{a\ln(G)+b}\right) = 0, \tag{3-17}$$

rewriting Eq. (3-17) in terms of the gain the following expression is obtained:

$$G_{threshold} = \exp\left(\frac{\ln\left(\frac{-1}{1+a}\right) - b}{a}\right), \tag{3-18}$$

Replacing this expression into the original equation, Eq. (3-7) we find the following intelligibility at which the ratio expressed in Eq. (3-14) reaches a maximum:

$$\hat{d}_{r\_thershold}(G_{threshold}) = \frac{100}{1 + \mathrm{e}^{a\ln(G_{threshold})+b}} = \frac{100}{1 - \left(\frac{1}{1+a}\right)}, \tag{3-19}$$

The maximum in the ratio (3-14) occurs around the turning point in the intelligibility where the intelligibility no longer rises linearly with the gain. This happens to occur around an intelligibility of 83-90%, with the defined gain range and the fact that it is expected that at the maximal gain the intelligibility is near to 100%. Through informal listening tests it was observed that these intelligibility percentages are satisfactory. To ensure good intelligibility the set-point is set at $d_{\text{set-point}}=0.65$ or equivalently $d_{\text{r\_set-point}}=88.6\%$. It should be clear that a set-point is chosen for the intelligibility and not for the gain. The calculated gain is specific to the noise level and the realizations of the speech and the noise.

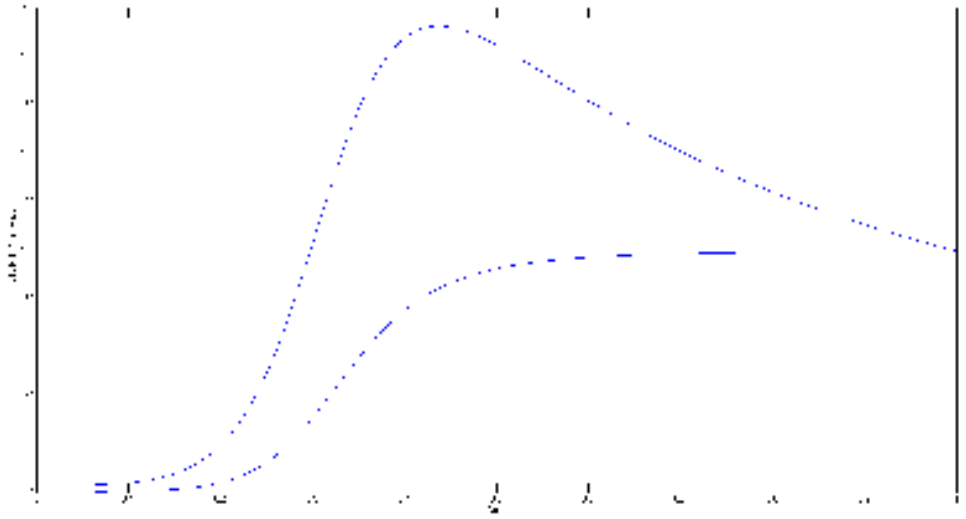

Figure 3-4 plot of ratio (3-14) (top) and the intelligibility Eq. (3-7) (bottom)

### 3.1.4. Optimal Controller

The optimization problem at hand is an optimization with a constraint. The constraint is the required intelligibility defined in the previous section. The cost function that is minimized could be defined by the following function:

$$\zeta = G^2, \tag{3-20}$$

Formally the optimization could be described as follows:

$$\arg\min(\zeta) = \arg\min(G^2) \text{ for } d(G) \geq 0.65, \tag{3-21}$$

The problem with this optimization is that the constraint $d(G) \geq 0.65$ is a non-linear function of not only the gain, but also the noise, and clean speech realization. Due to the definition of *Y,* Eq.(3-2), with the gain being caught in the square root it is not directly possible to express the constraint, as a gain constraint instead of an intelligibility constraint. However using the expressed relationship between the gain and intelligibility, which is monotonic in expectation it could be stated that there is one point on the gain curve where the constraint is met and the gain is minimal. Therefore the optimal controller can also be defined by a controller that makes sure the intelligibility is always maintained at the optimum intelligibility, the set-point. Is the intelligibility too low, then the constraint is not met, is the intelligibility too high then the gain is also too high.

An optimal controller can then be modelled by the system shown in Figure 3-5, where the error is calculated as being the difference between the set-point and the calculated intelligibility. The AVC then adapts the filter F in such a manner that the error is minimized, i.e. set to zero. In the case of only a gain control the filter F is a filter with only one coefficient, the gain factor.

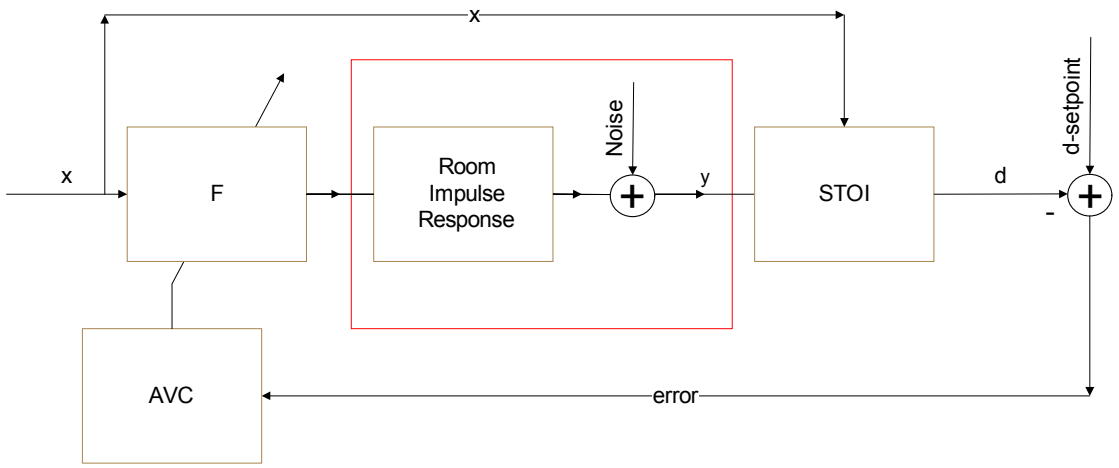


Figure 3-5 Optimal Controller

There is however an inherent problem with this controller that might be a reason for it not to work, guaranteed, in a real-life situation. That problem is caused by determinism, in the sense that the intelligibility of a message or frame cannot be calculated before the message has been played into the room and subsequently recorded by the sensing microphone. This is the result of, the exact noise-features not being known beforehand. Because an optimal controller is not possible, the next section (3.1.5) defines an error that can be used to see how far a designed solution is from this optimum. In section 3.3.3 the solution suggested by this thesis is presented, with its ways of staying close to the optimum.

### 3.1.5. Performance Evaluation of a Specific Solution

To verify the functioning of the proposed solution compared to other solutions it is interesting to define an error. This error should contain a penalty for the message not being intelligible and also, a penalty for being too loud. Using the optimal solution presented in the previous section, the error can be defined as the distance from this optimal point. To compare solutions, a fixed call with a specific fixed noise realization can be generated, and applied to both solutions. Then a Sum Square Error (*SSE*) can be calculated to define its performance.

$$SSE = \sum_{m} \left( d_s - d_m \right)^2 , \tag{3-22}$$

where $d_s$ is the set-point intelligibility ($d_s$=0.65) and $d_m$ the intelligibility of the $m^{\text{th}}$ frame. From this *SSE* the Mean Square Error (*MSE*) can also be calculated by dividing the *SSE* by the number of frames, *M*.

$$MSE = \frac{SSE}{M} , \tag{3-23}$$

There is however one problem with this error and that is, the relationship between gain, intelligibility and loudness is non-linear. Which has as a consequence that an error in unintelligibility, i.e. the intelligibility is below the threshold, weighs more heavily than the system being too loud, i.e. the intelligibility is above the threshold. This can be compensated for by adding extra weight to negative errors. In the proposed solution, this

imbalance is compensated for by designing an asymmetric controller that reacts differently on negative errors then on positive ones. The imbalance in the error could heuristically be compensated for by calculating the following two sum square errors separately for positive and for negative errors respectively:

$$SSE_{+} = \sum_{m} \left(d_s - d_m\right)^2 \text{ for } d_s - d_m > 0\,, \tag{3-24}$$

and

$$SSE_{-} = \sum_{m} \left( \frac{d_s}{1 - d_s} \right) \left(d_s - d_m\right)^2 \text{ for } d_s - d_m \leq 0\,, \tag{3-25}$$

The total sum square error can then be found by taking the sum of these two errors:

$$SSE = SSE_{+} + SSE_{-}\,, \tag{3-26}$$

## 3.2. REQUIREMENTS

Besides designing a controller that should as closely as possible match the optimal solution, presented in Section 3.1.4, the solution should meet a number of other requirements. These requirements originate from the nature of the PA-application. In appendix B a detailed description is given of all the requirements that where set including the specific requirements for the Bosch solution.

The first requirement is of course that the assignment is fulfilled; in designing a system that maintains the intelligibility and limits the loudness of the PA-system with respect to the ambient noise level, i.e. minimizes the error defined in Section 3.1.5. The remaining requirements can be split into two main sections. Firstly, under what conditions must the system be able to work? Secondly, how should the algorithm work and within which bounds?

### 3.2.1. Environmental requirements

The solution should work in any environment, where we can expect a PA system. For example a train station, airport, boat, shopping mall, office building, etc. These environments have very different acoustic properties, from small rooms with a lot of damping to very large spaces with many hard surfaces. Moreover the system should be able to handle changes in the room acoustics, for example more damping because of more people in the room or a new large object in the room, which changes echo paths, like for example, a train. Noise can be of any form or loudness. In the appendices, a further specification is made into the speeds of amplitude variations of the noise, and the spectral colour.

### 3.2.2. Algorithm requirements

The solution should be easy to use; this implies that the number of parameters that have to be set at a specific installation should be limited. Algorithmic complexity should be kept low so that it can be incorporated into the existing, or next generation PA systems. The proposed solution should not introduce a considerable delay into the signal path. The calculated gain of the controller should be within certain bounds, to limit the total difference between the loudest and softest level of the call. The system should react quickly to changing noise levels, like an incoming train. However the system should not react on impulsive noise, like shouting, hand clapping, or fireworks. Changes in volume should generally occur smoothly to hide the presence of the AVC to the public. Fast audible volume fluctuations can introduce new annoyances to the listening public.

During the absence of calls the algorithm should also adapt the gain of the amplifier, so that once a message starts, it starts at an appropriate volume. When operating with BGM, the volume fluctuations should be limited. The system should then only follow the general trend of the ambient noise. The system for example does not have to react on the incoming train. This limitation is emplaced because the audibility of the music is not as important as the intelligibility of a call.

### 3.3. PROPOSED SOLUTION

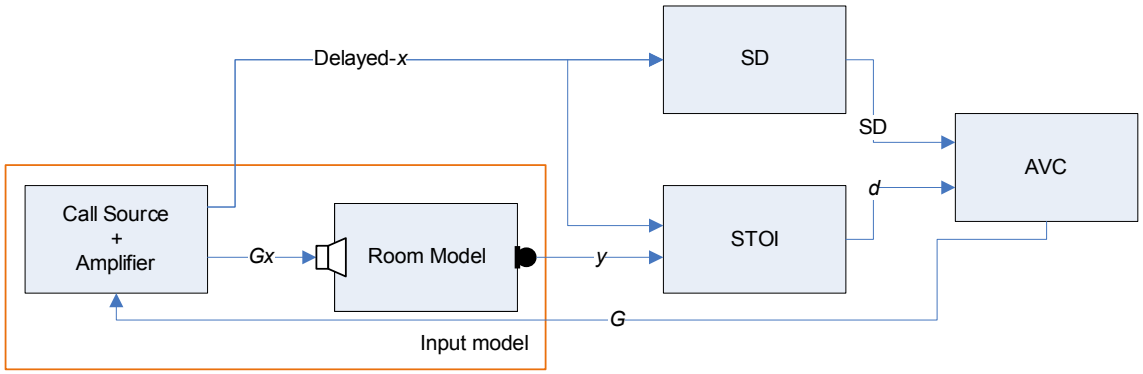


Figure 3-6 Basic control Model

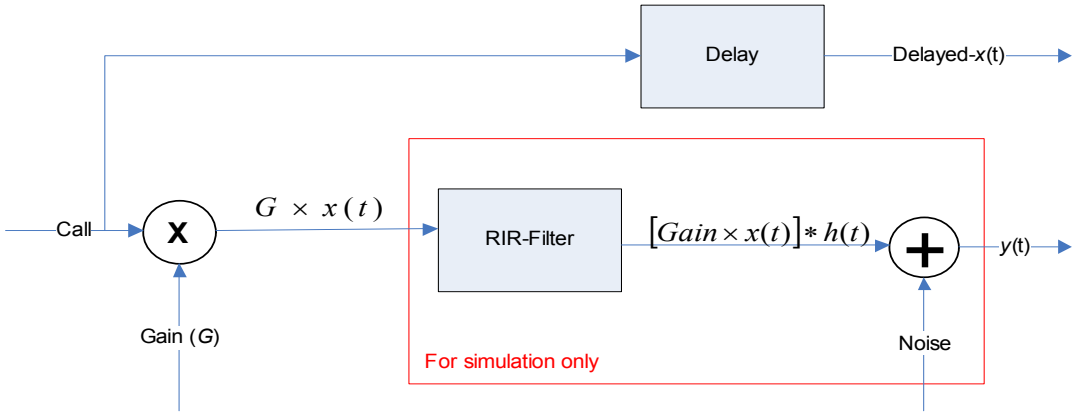


Figure 3-7 Input model

Working with the results from the theoretical analysis within the boundaries defined in the requirements section, a solution is proposed that can regulate the volume of a PA system. This section describes the overall structure of the proposed solution. In the next section, (3.4) a description is given on how this solution was modelled in Simulink for simulation purposes. Figure 3-6 shows the basic control model where the following blocks where designed; input model, signal detection (SD), AVC and a real-time implementation of the STOI algorithm. In Figure 3-7 a detailed input model is shown. The block that is specific for the proposed solution is the delay block. In the system, the 'simulation only' part models the characteristics of the room. The multiplier is a variable amplifier in the PA system chain. It is a simplification of the filter F shown in the optimal solution Figure 3-5. The delay block has as function to time align the signal *x* and *y*. This implies that the delay that is set should be equal to the delay from the amplifier to the sensing microphone. This delay can be automatically found or adaptively updated by finding the maximum correlation between the signal *x* and *y* (in relative silence). It is not expected that this delay varies considerably during normal operation.

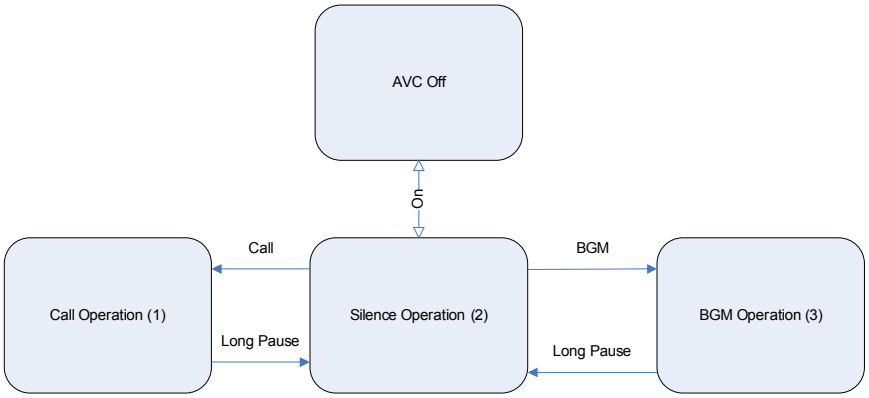

Figure 3-8 State diagram of the controller

As mentioned before the proposed solution should be able to handle both calls and BGM. It was also required in the previous section that the system should continue to adapt the volume in-between calls and BGM. The overall solution can be described by a state diagram (Figure 3-8) in which the system reacts differently depending on what state it is in. The thesis will focus on AVC during the call. AVC during BGM is, as for now, not fully implemented due to the lack of knowledge on the exact relationship between the calculated correlations using STOI, and the audibility of music. The operation during silences is strongly related to the simplest existing solution discussed in Section 2.1. In Appendix B a detailed description is given on how the system should operate in different states, and how and when the transitions occur. In a final implementation extra state transitions might be added, to accommodate for extra options, like a quick interruption of the BGM by a call, without having to go through the silence mode.

### 3.3.1. STOI adaptations

The STOI algorithm was in first instance not designed for real-time application, but the general structure of the algorithm does give room for a real-time implementation. The main factor that makes this possible is that the algorithm analyses the intermediate correlation coefficient over a certain time frame of approximately 400ms.
The following adaptations where made:

- Increased sampling frequency from 10000Hz to 11025Hz to allow for 4x down sampling from the standard audio sampling frequency of 44100Hz.
- Intelligibility analysis performed on a 384 ms frame basis with a frame-by-frame shift as discussed in the STOI model. I.e. the intermediate intelligibility, Eq. (2-5), is directly used and averaging over time is left out. This does however result in an estimate with a high variance.
- Signal detection is taken out of the STOI algorithm, the original algorithm removed all the silent frames from the signals to be analysed before calculating the intelligibility. The system now discards the information on intelligibility of the silent frames.
- As far as possible the algorithm was structured in such a way that the parallelism was exploited, by analysing all the one-third octave bands separately and simultaneously.
- Where possible the algorithm was implemented in such a way that typical DSP instructions like Multiply Accumulate could be used.

### 3.3.2. SD

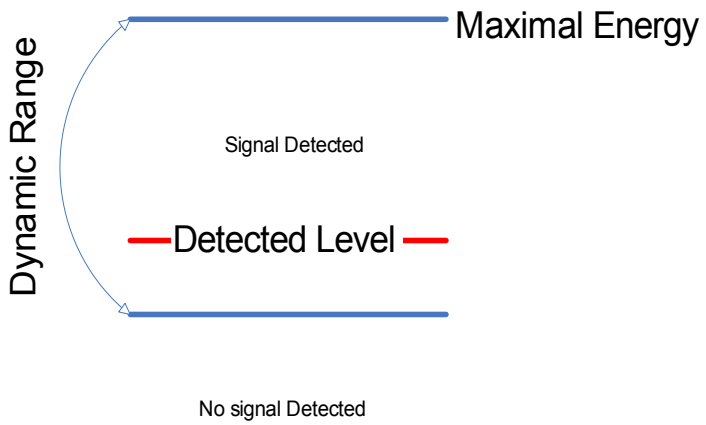


Figure 3-9 Signal Detection

It only makes sense to calculate the intelligibility of a frame if there is signal presence in the frame played out in the room. This is why signal detection is implemented. For signal detection, the clean signal is analysed. The analysis is performed on the same (400ms) frame, for which the intelligibility is calculated, i.e. the same 3840 samples of the clean speech. Within that frame the signal energy is calculated. The algorithm also keeps track of the maximum frame energy over time. The expected range in this energy level is predefined. Signal detection is then the result of the following equation:

$$SD = \{E_s - \max(E_s) + E_{range} \geq 0\},$$

where $E_s$ is the signal energy of the current frame, $E_{range}$ is a constant that fixes the expected dynamic range of this energy within speech. The maximum is tracked over time so that this maximum can adapt to different maximum frame energies, depending on for example the person who is doing the call. Figure 3-9 illustrates these boundaries and in which frame energy regions, signal is said to be detected and in which not. How this signal detection is used in the actual volume control is defined in the following sections 3.3.3 and 3.3.4.

### 3.3.3. AVC

The core of the proposed solution is the automatic gain controller itself. The AVC was designed as a proportional-integral-derivative controller (PID [18]). The choice was made for a PID controller because of its wide use and tune-ability. A PID controller is suitable this situation, because there is no forehand knowledge on the behaviour of the distorting noise and limited knowledge of the call signal. According to [18] a PID controller is the optimal controller if there is no knowledge of the underlying process. In this case there is some knowledge of the underlying process. And one of the recognized problems with the standard PID controller is that it has a symmetric response ([15]). This symmetry causes problems because the relation between our calculated error and the control parameter is not a linear one, as is shown in section 3.1.2. The problem would for example occur if the intelligibility were high; the system would relatively take a too small step back to keep the loudness limited.

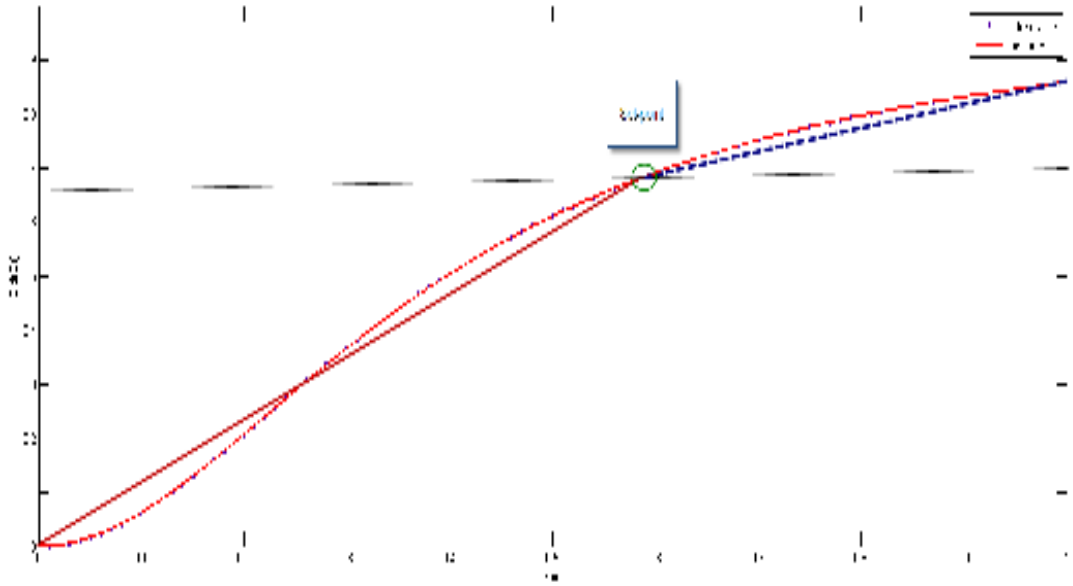

Figure 3-10 Asymmetric proportional control

To alleviate this problem an asymmetric PID controller is designed, that has a different behaviour for a negative and a positive error. If we start by analysing an asymmetric Proportional controller (P-controller), the applied gain would be proportional to the error. The error is the difference between the set-point, $d_s$, and the calculated intelligibility $d$.

$$error = d_s - d \ ,$$

If this asymmetric P-controller placed into the discovered mean relationship, between gain and intelligibility, it would result in a linear control across two gradients. One above the set-point, and one below. This illustrated in Figure 3-10. It is again important to mention that the depicted gains are for a specific noise level that was chosen for the abstract experiments. So there is not one optimal gain that can be selected throughout the operation. Using these gradients, it can be calculated how much the gain needs to be increased to reach the set-point intelligibility. This would however only work well if the intelligibility is not highly related to the specific noise and call realization. The integral(I) term adds extra control to a P-controller, the I-term keeps track of the previous errors, i.e., if the error is negative for a longer period of time the gain is decreased more because the integrator accumulates the errors. The applied gain is the sum of the terms. The variables $P_1$, $I_1$ and $P_2$, $I_2$ determine the weights of the proportional integral part for positive or negative error respectively.

Because the calculated $d$ and thus the error are noisy, the derivative term was set to zero. Resulting in the following asymmetric-PI controller, shown in Figure 3-11:

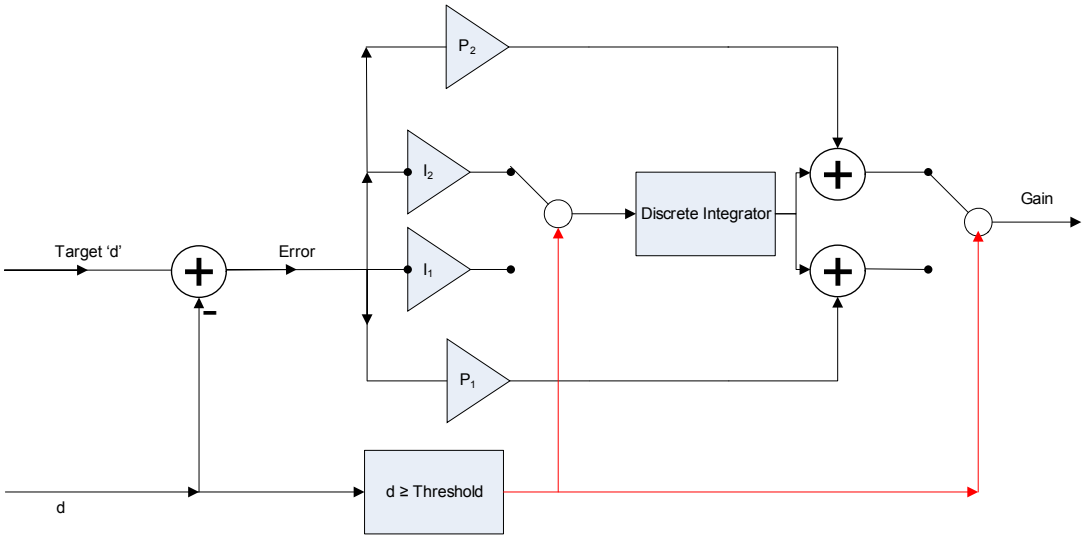

Figure 3-11 Asymmetric PI-controller for AVC

In Figure 3-11 the red line is a control line that activates the switch. If the error is negative, i.e. the call is intelligible, the proportional term $P_2$ and the integral term $I_2$ control the gain. In other case $P_1$ and $I_1$ are used.
In a PI-controller the choice of the parameters $P$ and $I$ is important. and determines the behaviour of the controller. The proportional term $P_1$ and $P_2$ are kept small because it is undesirable for the system to continuously adapt to instantaneous changes in intelligibility, which can be classified as $d$-estimation noise. The integral terms ($I_1$ and $I_2$) are the most important in this controller because they are less sensitive to noise in the error, due to the integration which is like an averaging over time. Secondly, the integral terms say something about the trend of the system, and reduce the time of unintelligibility, by increasing the gain more and more if the message remains unintelligible for a longer period of time. In the Simulink implementation, described in Section 3.4, the range of the integrator is limited for the system not to get stuck, at a gain that is too high or too low. As mentioned before the derivative term is set to zero because of the high variations in the error, these variations in the error would cause the derivative to be very high and highly fluctuating. If a smoother estimate of the error, or $d$, would be determined the derivative term could maybe be added to allow for more control.

Because a smooth gain adaptation is desirable for pleasant listening the gain is smoothed before it is applied in the amplifier. This does sometimes limit the speed and accuracy of control, but this is the trade-off between reaching the set goals, and designing a controller that is pleasant to listen to.

Signal Detection, as described in the previous section, is used to freeze the gain if there was no signal detected in the active frame. Moreover the state of the PI-controller is frozen so that the integrator is not contaminated by large invalid errors. If no signal is detected for a number of successive frames then the system switches to the silence mode which is described in more detail in section 3.3.4. The length of the echoes and reverberations determines the number of successive frames that have to be silent before switching to the other state. This can for example be the $RT_{60}$ of the environment translated to the equivalent number of frames. The $RT_{60}$ can vary from a few milliseconds to a number of seconds at for example a train station.

### 3.3.4. Silence mode

As mentioned in the previous section the system switches to the silence mode when a long pause is detected. The goal of this state is to maintain the gain level at appropriate level for when the message restarts. Because this state is only entered when there is no call signal, or remainder thereof, in the room, the sensing microphone only captures the ambient noise when in this state.

The ambient noise is sampled in a similar fashion, as the signal in the signal detector. The energy of the ambient noise is calculated for every frame. Using the maximum signal energy, from the SD-block, multiplied by the current gain the following hypothetical SNR is calculated;

$$SNR = \frac{\max(E_s)G}{E_{ambient}},$$

This SNR is then compared to a threshold. If the SNR is lower than the threshold the gain is increased if it is higher, the gain is decreased. A region around the threshold is also defined in which the gain is not adapted. This is done to not react on very small changes in the ambient noise. Moreover the increase/decrease is done in very small steps to only track the longer-term changes in ambient noise level. After the gain step is applied the new gain is clipped if necessary to the predefined gain range and then the value is stored.

### 3.3.5. BGM

The third mode of the algorithm is the BGM mode. This mode however has not been studied in depth so the control in this mode is also rather basic. It was seen through experiments that the 'intelligibility' can also be calculated for music. From now on intelligibility for music will be called audibility. The audibility is calculated with the same STOI algorithm. It was observed however that the calculated $d$ is generally lower, while the music was still audible. The relation between the calculated $d$ and the audibility deserves a deeper study and this is also suggested in the final section (4.1) of this Thesis.

As stated in the requirements, the volume should not adapt as rapidly during music as during speech. The BGM AVC uses the same controller as the proposed solution. The two factors that are changed are the threshold for control ($d\approx0.4$ instead of $d\approx0.65$), and the smoothing over time of the applied gain. SD and silence mode are also still active during BGM operation. Attention should be paid to how the transition from BGM to normal operation occurs, as for now this transition is required to go through the silence mode, with the echo/reverberation die out time included. The characteristics of music are however different from those of speech and this should be taken into account when looking at the effectiveness of BGM control. For example it can be disputed whether using such a low sampling rate is good enough for music. The rhythm, the clearest temporal feature in music is however contained in the current limited frequency band (Due to the limited sampling frequency). It was observed through tests that the calculated audibility is higher for music with clear rhythmic features, compared to music with a strongly varying or unclear rhythm.

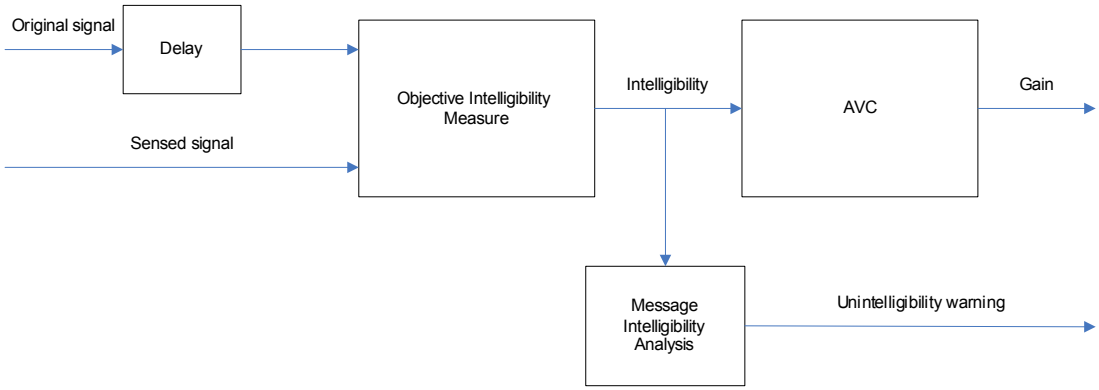


Figure 3-12 Possible extension using intelligibility based AVC

Maybe adding one extra band with the high frequency information would already be enough to capture the extra information needed, and to say something about the audibility of the BGM in total, not only the rhythmic features.

### 3.3.6. Possible extensions

With this new approach to regulating the volume of a PA system, a few extra option come within range of implementation. Because the intelligibility of the signal is continuously monitored, the general intelligibility of the whole call/message can also be calculated. If this intelligibility was not satisfactory in general, or a part longer than for example a word, the message could be repeated automatically. Figure 3-12 illustrates where this extra block could fit into the whole AVC chain. Feedback could be given to the call operator to repeat his/her message. Real-time information about the intelligibility could also be given to the call operator who could use this information to delay his/her message or repeat only the unintelligible part.

## 3.4. SIMULATION MODEL

To test the developed theory and algorithm a simulation model is made. This simulation model is developed in Simulink [16]. The simulation model is a direct implementation of the proposed solution. This section therefore only mentions the things that where specific to the design in Simulink, or the points that are relevant for understanding the tests described in the next chapter. The choice made for Simulink for the following three reasons:

1. Real-time simulation possible
2. Extensive signal processing block-set with audio I/O
3. Insight into possible hardware structure and possibility of porting to an FPGA.

To be able to simulate without using a real room a simple PA system model is designed that can simulate a real environment. The simulation model uses the input model shown in Figure 3-7. The RIR is acquired through a small Matlab-function that calculates a room impulse response from the dimensions of the room, location of the source, location of the microphone and the reflection coefficient of the surfaces. The delay is automatically extracted from the location of the maximum value of this impulse response. The speech and noise signal are both generated beforehand and normalized at -3dB. In Chapter 3.6 a detailed description is given of the specific noise samples generated.

### 3.4.1. Live set-up simulation model

For a live set-up the simulation only components are taken out of the control model. This means that the room model and the integrated noise source are removed. The input, *y,* is now the microphone and the output, a loudspeaker. The factor that has to be correctly set, is the delay between the microphone and the loudspeaker. Including the internal delays in the set-up, caused by the buffering of in- and output. To be able to simulate some large noise sources that cannot be brought into the test space, an extra sound source is used to play noises, like a passing train or a very large crowd.
The following hardware configuration is used in the tests:

- Desktop PC
- RME Hammerfall DSP multiface II
- Microphone
- Loudspeakers
- Microphone pre-amp
- Loudspeaker amplifier
- Noise source
- Matlab/Simulink (R2007b)

These components are linked together to form the following system depicted in Figure 3-13. The desktop PC runs the Simulink model.

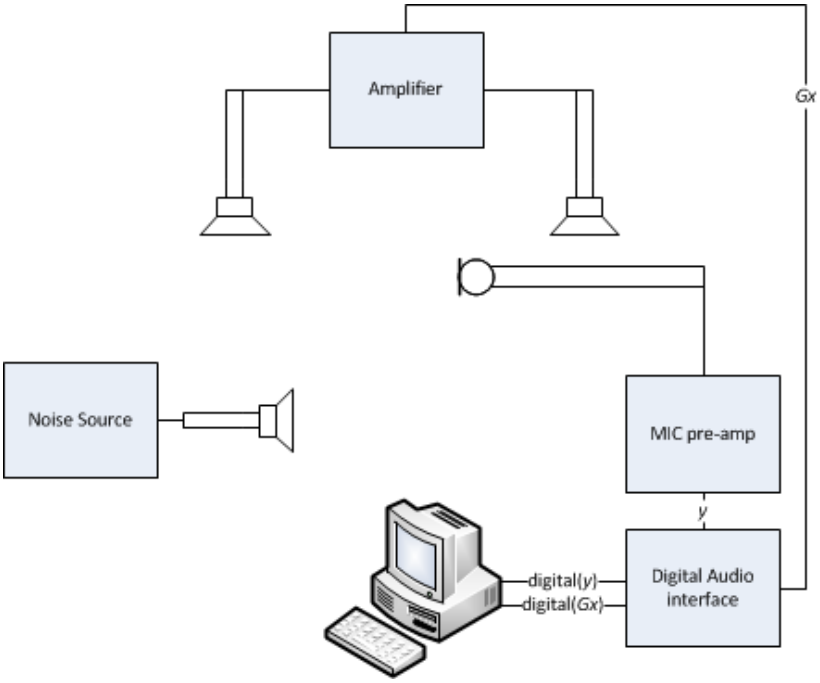

Figure 3-13 Live set-up

## 3.5. TESTS

To verify the correct functioning of the system, a number of tests were conducted. Two main sets of tests were executed. Initially tests where done with the simulation model (Section 3.6). Later a live set-up was designed that could test the algorithm in a more realistic manner (Section 3.7). The tests are designed to show that system requirements (Section 3.2) are met and to discover where there is still room for improvement.

One existing solution, the non-dynamic AVC (nd-AVC), is modelled using the silence mode designed for the proposed solution. When de proposed solution switches to silence mode, the nd-AVC adapts in the same manner as the proposed AVC in silence mode. When a call restarts the nd-AVC freezes its gain. This model is chosen as a reference model because of its easy implementation, and because it resembles the AVC currently being used by Bosch, where this thesis research was conducted.

This section contains the description of the tests that were performed (Section 3.7 and 3.8) followed by the results of the tests, and the discussion thereof, in Section 3.8. The more detailed description of the tests and the specific parameters that were used can be found in Appendix C.

## 3.6. SIMULATION TESTS

The simulation tests are designed in such a manner that all the extreme cases are tested. But also the general behaviour in for example stationary noise is tested. Then through visual inspection it is verified that the gain follows the noise level. Moreover the controlled announcement is also listened to, to observe whether it is intelligible and not too loud(on a Beyer Dynamic DT990 headphone). The performance is verified by calculating the weighted error, Eq.(3-26), defined in section 3.1.5.
Four factors are varied to generate representative tests.

1. The call
2. The room impulse response
3. The noise characteristics
4. The temporal level variations of the noise

For the call, there are three options; Speech without pauses, speech with pauses and background music. The first type of call is designed to make sure only the normal AVC mode is tested and not the silence mode. The second is more realistic and really tests the complete algorithm. The last is the BGM call which is used as an indication on how the system will react during BGM. The room impulse response is chosen to be for a small, medium or large space. The sizes chosen can be found in Appendix C.I. The noises are chosen to cover an as broad as possible range. The different options are: White noise, speech shaped noise (speech babble), pink noise and realistic noises like a passing train. The following temporal variations in level are analysed: No variations, slow variations, fast variations, instantaneous variations and pulses.

If all possible combinations would be made this would result in 180 tests. Some tests are however combined, by for example performing various level transitions consecutively in one test. Moreover for the realistic noises the original level variations are preserved. The BGM mode is also not tested in such great detail. By making different

combinations a test set is derived, consisting of 27 different tests per type of call. 9 noise files are made, which can be combined with a call or BGM file to perform the test. Each test is 60 seconds long. Both the noise and the call are normalized at –3dB.

### 3.7. LIVE TESTS

Initially the live-tests were performed using the same test set as defined for the simulation tests. The same call, BGM and noise files where used, with the difference that the noise was now mixed with the call in the room and not inside the model. The microphone now being present in the room also reacted on people talking inside the room. The live tests where useful to get a better insight into the loudness perception of the calls. Moreover, in a larger scale experiment the annoyance of many volume fluctuations becomes clearer. For a clearer picture however more live tests are needed using different loudspeaker configurations in different rooms.

### 3.8. TEST RESULTS

For every test the adjusted and the not adjusted audio (only with simulation) with the noise was stored. The calculated intelligibility and the applied gain were also stored for possible further analysis.
In analysing whether a test had passed the following four things where analysed:

1. Is the message intelligible (listening)
2. Is the message not to loud (listening)
3. Does the gain level follow the noise level in general and not vary too much (listening and visual inspection)
4. Calculate the error defined in section 3.1.5

#### 3.8.1. Simulation

Through the simulation test it was observed that the proposed solution outperformed the current solution and the SNR-solution. In Table 2 we can see the mean MSE over the various test performed. The Max and Min test, are the results for applying the maximal gain (0dB) and the minimum gain (-12dB) respectively. The old solution was modelled as described before by the silence mode of the proposed solution. The SNR system is a general model that represents an SNR-solution that tries to maintain the SNR at a certain value. The clean noise is used as the noise signal, representing a perfect filtering operation. The SNR is then calculated on the same frame and rate as the proposed model, allowing the gain to adapt accordingly.

| Test Set | Proposed | Old | SNR | Max | Min |
|---|---|---|---|---|---|
| 1 | 0.0153 | No results | 0.0212 | 0.0300 | 0.0322 |
| 2 | 0.0263 | 0.0361 | 0.0391 | 0.0374 | 0.0298 |

**Table 2 Test Results MSE**

For the first test set there were no valid results for the old solution because there were no silences in the speech to allow the algorithm to adapt. The mean square error for the second set of tests was calculated by dividing by the number of frames that contained a

signal. The silent frames were set to contain no error, and were thus not taken into account when calculating the error.

The differences between the different tests for the proposed solution show that the silence mode could be further improved to give the algorithm a better starting point after a silence. The difference in test 1 and 2, for Min and Max, could be explained by the different lengths of the signals containing speech and whether these where in noisy or silent parts of the total signal.

Observing the gain and the noise level as waveforms, showed that the proposed solution adapted the volume according to the noise. Listening tests supported that the message remained intelligible. Because the loudness could not be properly evaluated on the headphones, no valuable statements can be made about that part of the problem. In the live tests however the loudness aspect became clear. Moreover the test with pulses also showed that the set requirements where met.

It is clear from these results that the proposed solution outperforms all the other solutions. The SNR-solution could maybe be improved by analysing the SNR more specifically in speech regions, or over another frame length. This is supported by the test results that showed that especially for the Speech Shaped noise, the SNR-solution performed very badly. The improvement could however be undone by introducing a realistic signal separation filter, with its shortcomings.

### 3.8.2. Live Tests

As mentioned before the live tests where only basically conducted. What however became clear from the live tests was that controlling the loudness might require some more attention and modelling. Moreover it was observed that fast volume fluctuations are especially annoying when decreasing the volume, the volume control thus requires a smoother decay curve then what is now implemented. The trade-off between fast control and smooth/pleasant volume fluctuations also became evident through the live tests.

# 4. Conclusions

The proposed solution is an innovative approach to regulating the gain of a PA system. It has a strong focus on intelligibility and uses a direct approach to regulating the gain without having to filter the microphone signal. This specific focus ensures fulfilling the primary target of a PA system, 'getting the message across'. Moreover it opens the way to new innovations that can further improve this target, as discussed in Section 3.3.6. The proposed solution makes use of the available information contained in the signals to a full extent, opposed to other solutions, which only use the energy of the clean signal and the noise. The energy alone, of the call, is not the primary information carrier.

The choice to control the intelligibility around a certain optimum helped to reduce the loudness as well because of the relationship between gain and intelligibility (Section 3.1.2), and the relationship between loudness and the gain. Using a more accurate model, to quantize the loudness in noise could maybe improve the loudness reduction of the PA system. The risk is however, that for such a measure the noise, or noise level, has to be available as a separate signal. Separating the noise signal from the reverberations is a costly procedure that was intentionally avoided for this solution.

The calculated intelligibility per frame is rather noisy; this limits the possibility for a very fine gain control. The trade-off exists between desiring a quick reaction on the one hand and a smooth control on the other hand. Trying to acquire some forehand knowledge on the expected intelligibility of a specific fragment of a call, could also improve the smoothness of control. Such an option would however introduce a small delay into the call chain (call source – amplifier –loudspeaker – listener).

Using the algorithm to control for background music works, but is based on the heuristic assumption that the STOI algorithm also gives information about the audibility of music.

Another important issue which should be taken into consideration and requires more research is the placing of the sensing microphone. Because the interest goes out to the observation of the listener, in both the intelligibility and the loudness, it is preferred that the microphone is placed at the location of the listener's ear. This is rather impractical, let alone infeasible if the intelligibility and loudness have to be observed for a large group of listeners. If the conclusion of such research would be that the intelligibility in one location is not related to the intelligibility some distance away, the proposed method of control would not be a good method for controlling the gain. Observations during the live-tests however do not indicate that this is the case.

The insensibility to the acoustics of the room is the major advantage of this solution. The acoustic properties of the room could also change at any time without influencing the functioning of the AVC. Using an adaptive filter as discussed in Section 2.1 is not so insensitive to the acoustics. Designing an adaptive filter for very large spaces, with a lot of echoes, is a lot harder than for a small room.

Overall the presented solution is a fast AVC with a low complexity. The AVC ensures that the calls of the PA system are intelligible but not too loud. Moreover PA systems can be extended to improve intelligibility even further, when the limits of the PA system are reached by for example automatically repeating the message. Simulation tests

show that the proposed solution outperforms all the present solutions as shown in Section 3.8.1.

## 4.1. SUGGESTIONS FOR FURTHER RESEARCH

A few suggestions for future research where already mentioned in the main body and conclusion of this thesis. This section mentions the required, and possible, research questions that could improve, an intelligibility based AVC for public address systems.
The following research questions should or could be answered:

- What are the exact effects of echoes and reverberations on the intelligibility, and how can this be taken into account in making, or analysing, an objective intelligibility model?
- Could the control be better when allowing frequency dependant gain adaptations? Would this allow for higher intelligibility with a lower loudness? The naturalness of the speech should then also be taken into account.
- What does the intelligibility measured at one location say about the intelligibility at another location some distance away? This is important for a proper positioning of the sensing microphone. Or if the difference is significant it should be taken into account.
- Could increasing the number of sensing microphones improve the representative sensing of intelligibility?
- How could the correlations calculated by STOI be used to express the audibility of music?

# 5. References


[1] Cees H. Taal; Richard C. Hendriks; Richard Heusdens; Jesper Jensen; "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech", IEEE Transactions on Audio, Speech and Language Processing, 2011

[2] Oosterom, JA.; "Literature Survey: Feedback Cancelation Using Adaptive Filters", 2010

[3] Roy D. Yates; David J. Goodman; "Probability and stochastic processes:a friendly introduction for electrical and computer engineers", John Wiley & Sons, 2005

[4] Monson H. Hayes; "Statistical Digital Signal Processing and Modeling", John Wiley & Sons, inc., 1996

[5] Cees H. Taal; Richard C. Hendriks; Richard Heusdens; Jesper Jensen; "An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction", Proc. Interspeech 2009, 2009, pp. 1947-1950

[6] F. Cavallini ; "Fitting a Logistic Curve to Data", College Mathematics Journal, volume 24; Issue 3, 1993, pp. 247-253

[7] Park HK.; Bradley JS.; "Evaluating signal-to-noise ratios, loudness, and related measures as indicators of airborne sound insulation.", Journal of the Acoustical Society of America 126, 2009, pp. 1219-1230

[8] Alan M. Richards; Herbert H. Lehman; "Most Comfortable Loudness for Pure Tones and Speech in the Presence of Masking Noise", Journal of Speech and Hearing Research Vol.18, 1975, pp. 498-505

[9] Ira M. Ventry; Robert W. Woods; Martha Rubin; Wathina Hill; "Most Comfortable Loudness for Pure Tones, Noise, and Speech", Journal of the Acoustical Society of America Vol. 49, 1971, pp. 1805-1813

[10] Nick Hoogland; "AVC: Target Specification", Bosch Security Systems B.V., 2004

[11] Stephane M. d'Alayer de Costemore d'Arc; "Automatic process of adjustment of the volume of sound reproduction", US Patent 5,530,761, 1996

[12] Frank P. Helms; "Automatic volume control to compensate for ambient noise variations", US Patent 5,666,426, 1996

[13] Robert Thomas Munson; Philip Hodgson; "Variable gain amplifier controlled by ambient noise level", US Patent 3,934,084, 1974

[14] Seved Ali Azizi; "Apparatus and method for noise-dependent adaptation of an acoustic useful signal", US Patent 6,628,788, 2001

[15] Baskys, A; Zlosnikas, V.; "Asymmetric PID controller", IEEE Industrial Electronics, IECON 2006 - 32nd Annual Conference on, 2006, pp. 219-223

[16] Mathworks; "Simulink", www.mathworks.com/products/simulink/, 2007

[17] Sauert, B.; Vary, P.; "Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments", Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, 2006, pp. 493-496

[18] Bennet, Stuart; "A History of control engineering", Peter Peregrinus Ltd., 1993, pp. 48-48

[19] Leyser Claus; Albert, Rolf "Circuit arrangement for the automatic adaptation of the volume of a loudspeaker to a disturbing noise level prevailing at the loudspeaker location", DE 3338413(A1), 1985

[20] Gore, D.N.; Kemp, M.J; Puddifoot, G.J.; Hayton, C.; "Audio-visual reproduction", US Patent 6,370,254 B1, 2002

[21] Yokohama, Y.K.; Zushi, T.C.; Yokohama, K.Y.; "Sound reproduction apparatus", US Patent 5,796,847, 1995

[22] Habets, E.A.P.; Gannot, S.; Cohen, I.; Sommen, P.; "Joint Dereverberation and Residual Echo Suppression of Speech Signals in Noisy Environments", Acoustics, Speech and Signal Processing, IEEE Transactions on , Volume: 16, Issue:8, 2008, pp. 1433-1451

# 6. Appendices

## A. MATHEMATICAL PROOFS

### A.I Derivation of derivative with respect to the Gain

Quotient rule derivative of Eq. (3-3).

$$\frac{\partial d}{\partial G} = \frac{\mathbf{C}^{\mathrm{T}} \frac{\partial \mathbf{Z}}{\partial G} \sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}} \sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}}}{(\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}})^2 \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2} - \frac{(\mathbf{C}^{\mathrm{T}}\mathbf{Z}) \sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}} \frac{\partial \sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}}{\partial G}}{(\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}})^2 \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2}, \tag{(6-1)}$$

$$= \frac{\mathbf{C}^{\mathrm{T}} \frac{\partial \mathbf{Z}}{\partial G} \sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}} \sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}}}{(\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}})^2 \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2} - \frac{(\mathbf{C}^{\mathrm{T}}\mathbf{Z}) \sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}} \frac{\partial \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}{\partial G}}{2(\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}})^2 \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2 \sqrt{\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}}, \tag{(6-2)}$$

$$= \frac{\mathbf{C}^{\mathrm{T}} \frac{\partial \mathbf{Z}}{\partial G} \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2 \sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}}}{(\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}})^2 \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2 \sqrt{\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}} - \frac{(\mathbf{C}^{\mathrm{T}}\mathbf{Z}) \sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}} \frac{\partial \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}{\partial G}}{2(\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}})^2 \left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2 \sqrt{\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}}, \tag{(6-3)}$$

Cancel out $\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}}$ on the right hand and take $\left(\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)^2 = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$

$$= \frac{2\mathbf{C}^{\mathrm{T}} \frac{\partial \mathbf{Z}}{\partial G} \mathbf{Z}^{\mathrm{T}}\mathbf{Z} - (\mathbf{C}^{\mathrm{T}}\mathbf{Z}) \frac{\partial \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}{\partial G}}{2\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}} \sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}} (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}, \tag{(6-4)}$$

Filling in $\frac{\partial \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}{\partial G} = 2\mathbf{Z} \frac{\partial (\mathbf{Z})}{\partial G}$ and cancelling out the factor 2 Eq. ((6-4) then becomes:

$$= \frac{\left(\mathbf{C}^{\mathrm{T}} \frac{\partial \mathbf{Z}}{\partial G}\right)\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right) - \left(\mathbf{C}^{\mathrm{T}}\mathbf{Z}\right)\left(\mathbf{Z}^{\mathrm{T}} \frac{\partial (\mathbf{Z})}{\partial G}\right)}{\sqrt{\mathbf{C}^{\mathrm{T}}\mathbf{C}} \sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}} (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}, \tag{(6-5)}$$

## B. REQUIREMENTS

### B.I Context

A Public Address (PA) system is used for conveying messages (Calls) or Background Music (BGM) to the public inside the building or venue. The main components in a PA system are the following:

1. Base station producing calls or BGM. Connected to the network.
2. An audio network for signal transportation
3. Boosters (amplifiers), connected to the network, that amplify the signal for a specific listening area (zone)
4. Loudspeakers connected to a booster
5. Sensing microphone for recording the audible signal in the zone
6. AVC connected to the network or located inside the booster

The AVC makes use of both the original signal, and the signal recorded by the sensing microphone.

Calls can be of different nature but with all calls it is important that they are intelligible for the listener. Different aspects influence the intelligibility of calls. The main degrading effect on the intelligibility of a call is the Background Noise (BGN) present in the listening area. This BGN is a summation of noise produced by various sources. The goal of the AVC is to adjust the volume of the PA system to compensate for this BGN, maintaining a certain level of intelligibility. This includes reducing the volume if the BGN is low to minimize inconvenience caused by the PA to unintended listeners (for example people living close to a train station).

### B.II AVC

The AVC system has three modes of operation:

1. Normal operation during Call
2. Operation between speech / BGM
3. Background music operation

Where the first and the third mode are the modes in which AVC is active. Mode 2 gives a starting point for the gain level when the system switches to mode 1 or mode 3. The system switches back to mode 2 when no voice or signal is detected on the input.

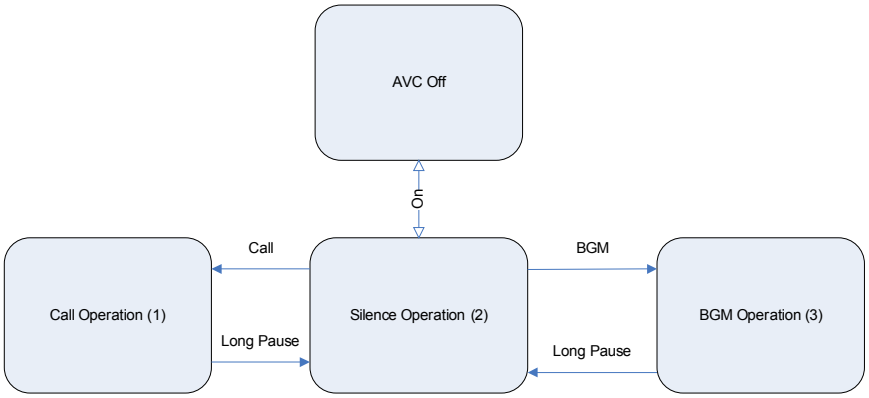

Figure 6-1 Mode of operation state diagram

### B.III Normal operation during Call

During normal operation the system should function as follows. With the limitations defined below. The system should react on noise sources that exist longer than the maximal impulse width. The system should adjust the volume in a pleasant manner. Volume changes should not be annoyingly audible to the general listener.

The system should not react on impulsive noise sources like:
- Fireworks
- Clapping
- Dropping of certain objects

An impulse width is limited to: 100 ms (smaller than the attack speed). STOI measure, *d*, should be maintained within: 0.6-0.7 range. It is of importance that for very low values of *d* that the gain adjustment is large to allow for quick reaction for quickly rising background noise to minimize the loss of intelligibility. The volume decay after BGN has died down should be smooth for a pleasant listing listening experience.

**Out of range**
If gain is 0dB (Maximal level) and STOI is below 0.4 for 3 or more consecutive frames the message was not or not completely intelligible and a warning must be given so that the system can repeat the message.

**Power**
Gain Range: -12dB – 0dB
Maximum signal Power: 100 dB SPL (dependant on the specific PA installation)

**Speed**
Attack speed: System should react within 200ms
Maximum attack: +3dB per frame
Attack curve: Dependent on STOI
Maximum decay speed: -2dB per frame
Decay curve (shape): logarithmic

### B.IV Operation between speech

Measure noise level, and determine what the input level should be with that noise level as a rough starting estimate for when the speech (re)starts. In this mode the adaptation rate should be lower (more smoothing) to not react on short-term changes in the BGN that will not affect the call dramatically when it starts.

**Power**
Gain Range: -12dB – 0dB
Maximum signal Power: 100 dB SPL (dependant on the specific PA installation)

**Speed**
Attack speed: System should react within 1000ms (for very smooth adaptation of the volume level)

### B.V Background music operation

The system should also adapt when Background Music (BGM) is played. The adaptation rate can however be slower and the smoothing over time is higher. The audibility measure should have another level. STOI should be maintained around the $d$=0.5. With the BGM music operation the system should always start at the lowest level when turned on and then adjust till the required level of audibility is achieved. This can be so because no important information for the listener is lost if part of the music is inaudible. During longer pauses in the music the system switches to mode 2.

**Power**
Gain Range: -12dB – 0dB
Maximum signal Power: 100 dB SPL (dependant on the specific PA installation)

**Speed**
Attack speed: System should react within 1000ms
Attack curve: log
Decay curve (shape): log

### B.VI Operation during Transitions

During transitions the system should function continuously. The lag in volume adjustment should be limited to maximally one frame, for the transition into mode 1 from mode 2. SD should correctly detect the speaker starting/stopping to talk within one frame. Signal detection should detect silences in music but should not switch on short pauses in the music or low-level parts of a song. During short pauses the system freezes the gain at the last calculated value. In mode 3 this freezing occurs for longer pauses up to. In both cases if no signal is detected for a longer period of time the system switches to mode 2. The system leaves mode 2 straightaway when a signal is detected. Signal detection is done on the same 400ms frame over which the intelligibility or audibility is calculated over.

### B.VII General requirements

The system must be stable and be fail-safe.
The system should not require a lot of calibration at installation.
Sampling frequency: 11025 Hz
Sampling frequency input signal: 44.1 kHz

**Environment**
Maximum noise Power: … dB SPL (to be defined)
The power of the echo: … dB (to be defined)
Length of the maximal echo/reverb: … ms (to be defined)

**Installation parameters**
The following parameters have to be set on installation
- Delay between microphone and primary loudspeaker
- Required intelligibility level (dependent on echoic nature of the room) (set during a test in silence)

**Location of the sensing Microphone**
The sensing microphone should record a signal that is representative to what the listener hears. So a representative mix of, the PA signal and the noise, as little influence as possible from very local noise sources.

## C. TEST DOCUMENT

This document defines the test that the AVC has to undergo to prove its robust functioning and verify that the requirements are met.

### C.I Simulation Setup:

The simulation will test the extreme cases under which the system must be able to operate. The simulations should be conducted in a Matlab/Simulink environment.

With the following test parameters/files for the PA system:
- 3 different room impulse models (small, medium, large)
- 3 different source samples in different languages
  - No silence call messages
  - Call messages with natural pauses
  - Music files with a classical music fragment and a segment of popular music.

With the following test parameter/files for the noise:
- 3 different average gain levels (soft, average, as loud as PA limit)
- Noise sources should cover the following range of conditions
  - Slow varying white background noise
  - Slow varying SSN (Speech Shaped Noise)
  - Impulsive noise
  - Fast varying background noise (white, SSN, limited frequency)
  - Step varying background noise

**Mode 1**
Play the three different calls (Female/male speaker / other language/ music) over all different noise sources and levels.

### C.II Test List:

**RIR (**W*L*H**):**
Small:
-Room size (5*10*2)
-Location source (x,y,z) (1,1,1)
-Location microphone (2.5,5,1)
-Reflection coefficient: 0.6

Medium:
-Room size (10*20*5)
-Location source (1,1,4)
-Location microphone (5,10,4)
-Reflection coefficient:0.6

Large:
-Room size (30*50*15)
-Location source (1,1,14)
-Location microphone (15,25,14)
-Reflection coefficient:0.6

| Test # | RIR | Speech Char. | Noise Char. | Sound file *.wav |
|---|---|---|---|---|
| 01 | Short | No Silence Speech(60s) -3dB | Noise 1 | Test01 |
| 02 | Short | No Silence Speech(60s) -3dB | Noise 2 | Test02 |
| 03 | Short | No Silence Speech(60s) -3dB | Noise 3 | Test03 |
| 04 | Short | No Silence Speech(60s) -3dB | Noise 4 | Test04 |
| 05 | Short | No Silence Speech(60s) -3dB | Noise 5 | Test05 |
| 06 | Short | No Silence Speech(60s) -3dB | Noise 6 | Test06 |
| 07 | Short | No Silence Speech(60s) -3dB | Noise 7 | Test07 |
| 08 | Short | No Silence Speech(60s) -3dB | Noise 8 | Test08 |
| 09 | Short | No Silence Speech(60s) -3dB | Noise 9 | Test09 |
| 10 | Medium | No Silence Speech(60s) -3dB | Noise 1 | Test01 |
| 11 | Medium | No Silence Speech(60s) -3dB | Noise 2 | Test02 |
| 12 | Medium | No Silence Speech(60s) -3dB | Noise 3 | Test03 |
| 13 | Medium | No Silence Speech(60s) -3dB | Noise 4 | Test04 |
| 14 | Medium | No Silence Speech(60s) -3dB | Noise 5 | Test05 |
| 15 | Medium | No Silence Speech(60s) -3dB | Noise 6 | Test06 |
| 16 | Medium | No Silence Speech(60s) -3dB | Noise 7 | Test07 |
| 17 | Medium | No Silence Speech(60s) -3dB | Noise 8 | Test08 |
| 18 | Medium | No Silence Speech(60s) -3dB | Noise 9 | Test09 |
| 19 | Long | No Silence Speech(60s) -3dB | Noise 1 | Test01 |
| 20 | Long | No Silence Speech(60s) -3dB | Noise 2 | Test02 |
| 21 | Long | No Silence Speech(60s) -3dB | Noise 3 | Test03 |
| 22 | Long | No Silence Speech(60s) -3dB | Noise 4 | Test04 |
| 23 | Long | No Silence Speech(60s) -3dB | Noise 5 | Test05 |
| 24 | Long | No Silence Speech(60s) -3dB | Noise 6 | Test06 |
| 25 | Long | No Silence Speech(60s) -3dB | Noise 7 | Test07 |
| 26 | Long | No Silence Speech(60s) -3dB | Noise 8 | Test08 |
| 27 | Long | No Silence Speech(60s) -3dB | Noise 9 | Test09 |

| Noise File | Noise Type | Gain characteristics |
|---|---|---|
| 1 | White | Stationary -3dB 60s |
| 2 | White | t=0 -33dB, t=30 -3dB, t=60 -33dB, linear |
| 3 | White | t=0 -33dB, t=10 -3dB, t=20 -33dB, t=25 -3dB, t=30 -33dB, t=35 -3 dB, t=40 -33dB, t=42 -3dB, t=44 -33dB, t=46 -3dB, t=48 -33dB,t=50 -3dB, t=50 -33dB, t=55 -33dB,t=55 -3dB, t=60 -3dB. |
| 4 | SSN | Stationary -3dB 60s |
| 5 | SSN | t=0 -33dB, t=30 -3dB, t=60 -33dB, linear |
| 6 | SSN | t=0 -33dB, t=10 -3dB, t=20 -33dB, t=25 -3dB, t=30 -33dB, t=35 -3 dB, t=40 -33dB, t=42 -3dB, t=44 -33dB, t=46 -3dB, t=48 -33dB,t=50 -3dB, t=50 -33dB, t=55 -33dB,t=55 -3dB, t=60 -3dB. |
| 7 | white | Impulses 1,10,20,50,100,300,400,600,1000ms -3dB each pulses repeated 2 times |
| 8 | Pink | Block wave form, 2s signal followed by 3 seconds -60dB silence. Blocks with the following amplitude -20,-15,-10,-8,-6,-4,-3,-6,-8,-10,-15,20 dB |
| 9 | Realistic | Passing Train, Stopping Train, Passing race car, Bar Panic. All normalized at -3dB with increasing gain. |

Similar tests are performed with silent speech test_2_01.wav to test_2_09.wav

**Mode 2**
When in mode 2 it is important that the system is sufficiently able to track the noise level. The most interesting part is the transition from mode 2 to one of the other two modes. It is important that mode 2 gives a good starting point for mode 1 or mode 3 for the gain adjustment.

**Mode 3**
Test for 2 different music segments, with all the pre-mentioned noise sources. One of the music files must be upbeat and contain many/large volume variations.

**Mode transitions:**
**Off – Mode 2:**
- Switch system on with low background noise level.
- Switch system on with high background noise level.

**Mode 2 – Mode 1:**
- Start a call with low background noise level
- Switch system on with high background noise level.

### C.III Live setup:

To test in a live situation a PA setup is needed that is as representative as possible as a standard installation, meeting al the set requirements. The live test should take place in two different spaces one representative of a smaller environment, with medium echoes, and a second in a large room with a lot of echoes. The second space should acoustically be comparable to a large train station or airport. For live simulations using a PC as a platform for the algorithm the following setup can be used:

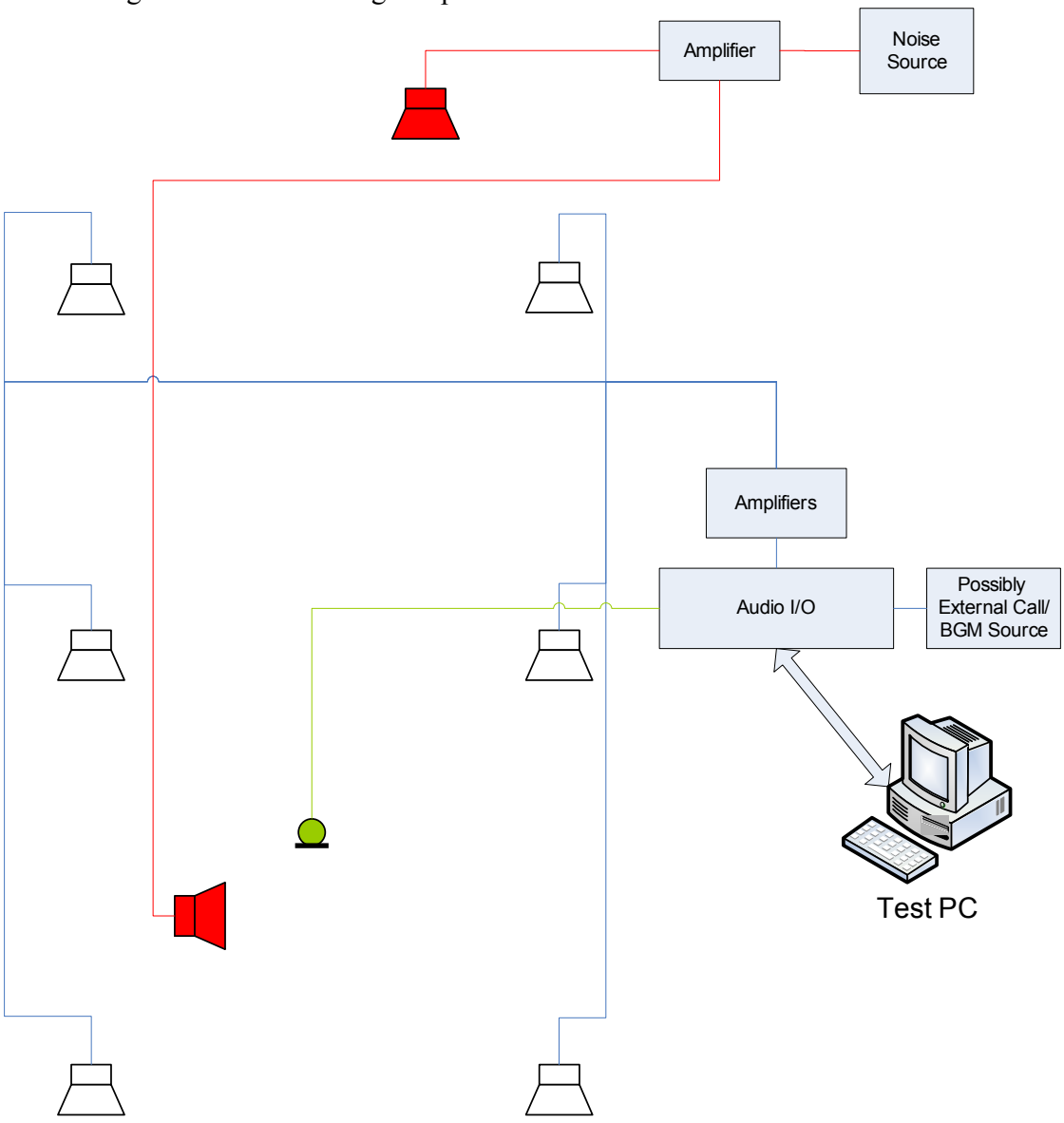


Figure 6-2 Setup for live tests

This system contains the following Elements:

- Test location (factory hall / transport hall)
- 6 PA Loudspeakers
- Mounts (stands) for temporarily mounting the Loudspeakers at an appropriate height within the test space.
- Amplifiers to drive these loudspeakers. (The amplifiers should directly connect to the audio I/O device).
- A noise source (CD-player / PC) with at least two channels so that multiple noises can be played at the same time through the different loudspeakers.
- Full range loudspeakers for the noise.
- Driving amplifier(s)
- Noise samples (similar to the samples defined for the simulation tests)
- Audio I/O device as interface for the test PC. At least 3 audio I/O channels.
- Omni-directional sensing microphone
- Microphone stand for mounting the sensing microphone in a desired position
- Possibly external Call and Background Music source.
- Fast test PC for running the designed algorithm (with Matlab, Simulink)
- Display for test PC
- Headphones for monitoring the PC
- Call + BGM samples similar samples to the ones used in the simulation tests.
- All the connecting cables
- Power source + cables

**Mode 1**

Play pre-recorded messages with all different pre-mentioned noise sources with the following variations:

- Vary the number of PA loudspeakers used
- Perform tests at different volume levels
- Move the sensing microphone around in-between tests and during a test.
- Change the positions of the noise sources with respect to the sensing microphone.

**Mode 3**

Test with 2 music files and noise track played on separate loudspeakers

### C.IV Test proposed for future work

- Implement algorithm on hardware platform (FPGA/DSP) and re-perform the live tests.
- Place the algorithm within the existing Bosch PA architecture/network
- Test at a real location with real noise sources.