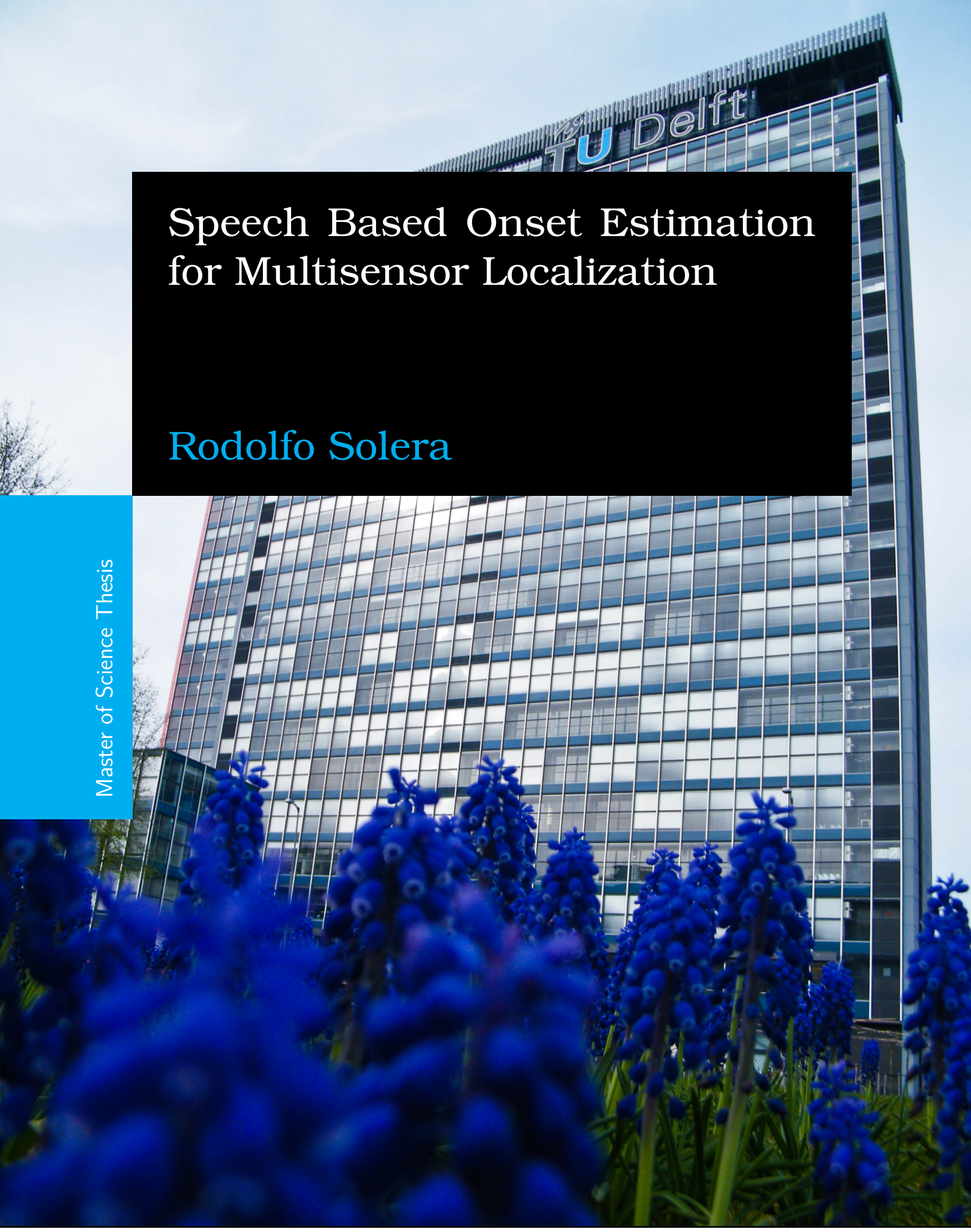



# Speech Based Onset Estimation for Multisensor Localization

Rodolfo Solera

Master of Science Thesis

TU Delft Delft University of Technology

Circuits and Systems

# Speech Based Onset Estimation for Multisensor Localization

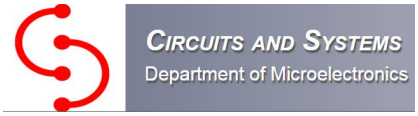
Master of Science Thesis

For the degree of Master of Science in Signals and Systems at Delft University of Technology

Rodolfo Solera

November 13, 2015

Faculty of Electrical Engineering, Mathematics and Computer Science(EEMCS) · Delft University of Technology

Delft University of Technology
Department of
Circuit and Systems (cas)

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science(EEMCS) for acceptance a thesis entitled

Speech Based Onset Estimation for Multisensor Localization

by

Rodolfo Solera

in partial fulfillment of the requirements for the degree of

Master of Science Signals and Systems

Dated: November 13, 2015

Supervisor(s):

prof.dr.ir. Richard Heusdens, Supervisor

Reader(s):

dr.ir. Jorge Martinez, Reader

prof.dr.ir. Gerard Janssen, External reader

# Abstract

*This work presents a study of a current problem in the field of audio processing: Source and receiver localization. Currently, this problem requires that either the onset time of the sources or the internal delay of the receivers are known. The algorithms studied here, take advantage of the structure of the time matrix, which contains the TOA of all the receivers with respect to all the sources, and finds the solution to the locations when the onset times are known. The problem here is then approached from a time difference of arrival (TDOA) perspective, which inherently cancels the onset times by subtracting the time of arrival (TOA) of a source at every receiver. The modification under the TDOA perspective however, proves not to be suitable under that framework. Therefore, a different approach is proposed, which uses speech signals as calibration signals in order to estimate the onset times. Such an approach is based on an algorithm which uses artificial calibration signals to calculate the onsets. Those signals are known a-priori, which implies that an additional device which produces those signals is needed. Once both internal delays and onset times are known, the locations of both sources and receivers can be estimated using a current algorithm which is also described here.*

# Table of Contents

# List of Figures

To my parents

# Chapter 1

# Introduction

In the field of multi-channel audio processing it is often required to determine the location of either sound sources $s$, sensors (microphones) $r$, or both (see an example of a layout in Figure 1-1. Once these relative locations are determined it is possible to carry out beamforming, and noise reduction techniques. Possible applications include conferencing systems and hearing aid systems, as an accurate estimation of the position of sources and receivers improves the quality of ad-hoc audio systems. Ad-hoc conferencing systems which use an array of smartphones (with no anchors) require the location of the sources and receivers to function correctly.

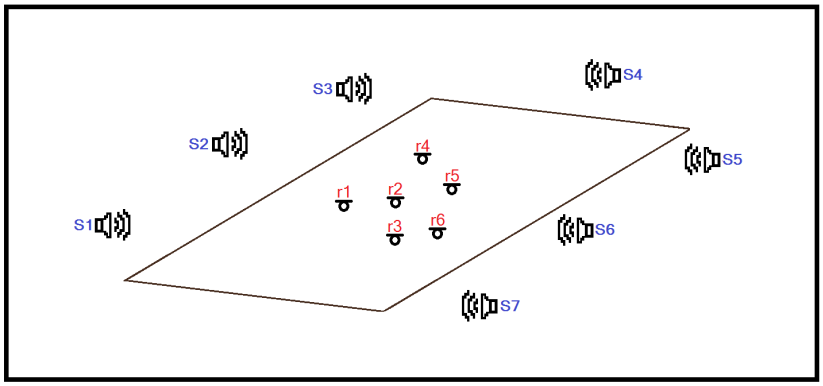



**Figure 1-1:** General layout. Each $r_i$ represents a receiver and each $s_j$ represents a source

There are many methods that can be used to find these locations when either the onset time $t_o$ or the internal delays $t_d$ are known (e.g. [2],[3],[4],[5]), or the relative distances between sources and receivers ([6]). This means that in order to determine the locations, certain information(such as onset times) has to be known. In order to estimate the locations it is first necessary to find the onset times and internal delays of the signal (figure 1-2). The onset time $t_o$ represents the instance in time at which the received signal is emitted. These internal delays $t_d$ are caused by the latency inherent to certain audio devices, such as devices that operate on Android or other operating systems (OS).

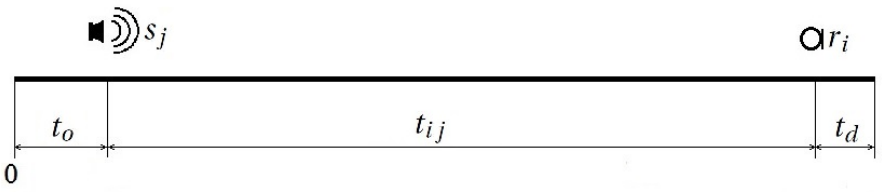



**Figure 1-2:** Decomposition of the measured TOA. Source $s_j$ emits at $t_o$, and $t_{ij}$ later, the signal arrives at receiver $r_i$, which has an internal delay $t_d$.

The time of arrival of a signal is related to the position of the receiver relative to the source of the signal, and this quantity is used in ranging calculations widely. In this case ranging is referred to as the finding of distances between a source $j$ and a receiver $i$. In general, finding the location of a receiver and a source can be laid down as a maximum likelihood problem which is expressed in (1-1). There $r_i$, $s_j$, $c$, $t_{ij}$, $n$ and $m$ represent $i-th$ receiver, $j-th$ source, the speed of sound, the estimated TOA, the number of sources, and the number of receivers, respectively.

$$\sum_{j}^{n}\sum_{i}^{m}(||r_i - s_j|| * c - t_{ij})^2. \tag{1-1}$$

The problem of finding the time of arrival ($||r_i - s_j|| * c$) has to be solved using the apparent received time of arrival, which will be called $ta_{ij}$ (see 1-2). The tag "apparent" of $ta$, comes from the fact that it is not equal to $t_{ij}$ as it also includes the onset time $t_o$ and internal delay $t_d$.

$$ta_{ij} = t_o + t_{ij} + t_d. \tag{1-2}$$

Therefore in order to find $t_{ij}$ it is necessary to first find the onset time $t_o$ and internal delay $t_d$. Several algorithms are used to find either one of those unknowns from the apparent time of arrival, but this does not solve the problem for jointly finding both $t_o$ and $t_d$. If there was a way to use the time difference of arrival (TDOA) between the receivers in an array (see $r_i$ as in Figure 1-1) then the $t_o$ would be eliminated. This approach is shown in (1-3), where $k$ and $i$ represent two different receivers and $j$ represents one source. The use of these TDOAs eliminate the need to determine the onset, and allow the finding of the internal delays $t_d$.

$$\begin{aligned} ta_{ij} - ta_{kj} &= (t_{oj} + t_{ij} + t_{di}) - (t_{oj} + t_{kj} + t_{dk}) \\ ta_{ij} - ta_{kj} &= (t_{ij} - t_{kj}) + (t_{di} - t_{dk}). \end{aligned} \tag{1-3}$$

The aim of this research work is to exploit the concept of time difference of arrival (TDOA) of calibration signals in order to determine the positions of receivers and sources of such signals. Current algorithms that aim at finding the internal delays are based on the TOA and require the knowledge of the onset times. So, if the effects of the onset time are eliminated from the calculation of the TOA, what is left is to find the internal delays $t_d$ of the receivers in order to then be able to determine the positions of receivers and sources. This is done for the time

of arrival case in [3] and [2], and then, if it was possible to replicate the structures exploited there, using the time difference of arrivals, it would be possible to find the localizations of receivers $r_i$ and sources $s_j$.

Another approach to receiver and source location is to use a pulse generator (e.g. a wavelet generator) with a known period ([7]). This approach uses the known pulse period of a calibration signal in order to determine the onset time $t_o$ (see section 3-0-3). These calibration signals are produced by an external source, for example a wavelet generator (a clicker). The known period of the signal from this external source is used to determine the onset times. Again, once the onset time is determined, the other unknowns can be found using known methods (e.g. [2], [4], [3]).

An external device such as the clicker mentioned before, represents an additional piece of equipment to be used in the localization process. This can be problematic for instance if such a device is not available. So as a way to improve on this approach in this project it is also researched if it is possible to arrive at the same results using arbitrary calibration sources instead of a pulse generator, for example speech signals. If this is possible and the self calibrating system does not depend on any external devices then the localization process would be more convenient.

# Chapter 2

# Main objective

The main objective of this work is to develop an algorithm that speech signals to facilitate the estimation of the locations of the sources an receivers.

## 2-1 Specific Objectives

To research current methods of localization of sound sources and receivers and identify the areas where they can be improved.

To improve current methods by modifying the framework to use time difference of arrival (TDOA), thus eliminating the need to estimate the onset times.

To develop an algorithm which makes use of speech signals instead of other synthetic control audio signals, to estimate the locations of the sources of those signals and the locations of the receivers.

To implement the algorithm in Matlab, simulate results and test it in a real life scenario for validation.

# Chapter 3

# Current algorithms

There are two different internal delay estimation algorithms studied in this research ([2] and [3]). Both have the purpose of finding the locations of sources and receivers by first estimating the internal delays $t_d$, when the onset times $t_o$ are known. As explained before, the apparent time of arrival $ta_{ij}$ includes an onset time $t_o$ as well as an internal delay $t_d$ (see equation (3-1)). Both the onset time and internal delay are unknown (the receiver gets $ta_{ij}$). Also, the internal delay is related to the latency at the receiver while the onset is related to the actual emission time at the source.

$$ta_{ij} = t_o + t_{ij} + t_d. \tag{3-1}$$

The algorithms described here, aim at finding the internal delays $t_d$ at each receiver (see figure 1-2). In the mentioned figure, $t_i$ is the start time of the sensor,$t_o$ is the onset time which represents the actual moment at which the actual signal was emitted, $TOA$ represent the actual time of arrival, and $t_d$ is the internal delay of the receiver.

The first algorithm ([3]) exploits the structure of the time difference relation to locations to find either the onset times or the internal delays in a closed form. The second algorithm ([2]), also exploits the structure of the TOA relationship but it uses a low rank approximation of the location matrix product ($\mathbf{R}^T\mathbf{S}$) with a recursive algorithm to find the internal delays.

Also, an autocalibration method is studied [7] and it exploits the use of an external calibration source. By using such a source of which the period is known, then it is possible to correctly determine the onset time $t_o$. After that, it is possible to find the internal delays $t_d$ using the methods in [2] or [3].

Finally, the last method studied [4], is used to determine the locations of receivers and sources, when the times of arrival are fully known. This means that in order to use this algorithm both the onset times $t_o$ and the internal delays $t_d$ have to be known in advance.

### 3-0-1 Pollefeys

The first algorithm to be analysed ([3]) has been developed assuming that the internal delays at the receivers are either zero or known values. Then the purpose of the algorithm is to find the onset time of the signal, regardless of its type.

A $(m \times n)$ time matrix $\mathbf{T}$ is formed from the $j = 1, ..., n$ and $i = 1, ..., m$ time elements from (3-2). This matrix is not square when the number of sources $(n)$ and the number of receivers $(m)$ is different. It contains the time of arrival of the sound to each of the microphones from each of the sources. In (3-2), $\hat{t_{ij}}$ represents the real time of arrivals between source $i$ and microphone $j$, and $t_{dj}$ represents the internal delay associated to the $i$ receiver.

$$t_{ij} = (\hat{t_{ij}} - t_{di})^2. \tag{3-2}$$

If $t_{di}$ is unknown, it is possible to find it if a periodical source of known period is used (see [7]). Though originally intended to find the time of departure (here, onset time), it is possible to find the internal delays instead. Also, the $R^T S$ vector product ((3-3)) contains the expressions which are equal to the norm of the distances between each microphone and source. This product is equal to $t_{ij}$.

$$R^T S = t_{ij}. \tag{3-3}$$

The first two elements of (3-3) are the position vectors which contain the individual positions of sources $S$ and receivers $R$ ((3-4)). Like mentioned above, the product of these, is equal to the norm of each of the dimensions of each of the distances between microphones and sources.

$$R = [(X_i^2 + Y_i^2 + Z_i^2)\ X_i\ Y_i\ Z_i\ 1] \text{ and } S = [1\ x_j\ y_j\ z_j\ (x_j^2 + y_j^2 + z_j^2)]. \tag{3-4}$$

The right hand side of (3-3) is expanded as ((3-5)):

$$v^2 * (t_{ij}^2 - 2t_{ij}t_{di} + t_{di}^2). \tag{3-5}$$

It is possible to rearrange the left hand side by creating a new $R$, indicated as $\tilde{R}$. This new term contains the $t_d^2 i$ term from the right hand side of the equation. Then (3-3) is modified as follows:

$$\begin{aligned} \tilde{R} &= [((X_i^2 + Y_i^2 + Z_i^2) - (v^2 * T_{di}^2))\ X_i\ Y_i\ Z_i\ 1] \quad , \\ S &= [1 \quad x_j \quad y_j \quad z_j \quad (x_j^2 + y_j^2 + z_j^2)]. \end{aligned} \tag{3-6}$$

Also from (3-6) it is possible to rearrange the expression as follows in (3-7). In (3-7), $\mathbf{D}$ is a diagonal matrix which contains the unknown $t_d i$ elements. Matrix $\mathbf{A}$ is formed by the $T_i^2 j$ elements and $\mathbf{B}$ is a matrix composed with the $-2t_i j$ coefficients. Matrix $\mathbf{T2}$ contains the $t_{ij}^2$ elements, $\mathbf{T}$ contains all $t_{ij}$. Matrices $\mathbf{S}$ and $\tilde{\mathbf{R}}$ contain the positions of all the $n$ sources and $m$ receivers respectively.

$$\begin{bmatrix}\mathbf{I} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} = (\mathbf{T2} - 2\mathbf{TD}) = \tilde{\mathbf{S}}^{\mathbf{T}}\mathbf{R}\,, \tag{3-7}$$

where matrices $\tilde{\mathbf{R}}$ and $\mathbf{S}$ are formed by vectors $\tilde{R}$ and $S$ from (3-6) as in the following expressions:

$$\begin{aligned}
\tilde{\mathbf{R}} &= \begin{bmatrix} ((X_1^2+Y_1^2+Z_1^2)-(v^2*T_{d1}^2)) & X_1 & Y_1 & Z_1 & 1 \\ & & \vdots & & \\ ((X_i^2+Y_i^2+Z_i^2)-(v^2*T_{di}^2)) & X_i & Y_i & Z_i & 1 \end{bmatrix}, \\
\mathbf{S} &= \begin{bmatrix} 1 & & 1 \\ x_1 & & x_j \\ y_1 & \cdots & y_j \\ z_1 & & z_j \\ (x_1^2+y_1^2+z_1^2) & & (x_j^2+y_j^2+z_j^2) \end{bmatrix}, \\
\tilde{\mathbf{R}}\mathbf{S} &= \mathbf{T2} - 2\mathbf{DT}.
\end{aligned} \tag{3-8}$$

Because the first row of matrix $\mathbf{S}$ is made of ones, there should be a linear combination of at least 5 rows of $\tilde{\mathbf{R}}^{\mathbf{T}}\mathbf{S}$, which results in a row of ones. Thus, there must be a vector $\mathbf{C}$ which fulfils ((3-9)):

$$C\begin{bmatrix}\mathbf{I} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} = [1\ 1 \cdots 1]. \tag{3-9}$$

From here it is possible to find $[\mathbf{CICD}] = X$ and vector $X$ can be split into two parts. Taking the second half $\mathbf{CD}$, and performing a piecewise division with the first half $\mathbf{CI}$, the result of this division is one of the elements of $\mathbf{D}$. This is possible because the size of $\mathbf{I}$ is known:

$$\begin{aligned} t_{di} &= X_{k+m}/X_k \\ k &= 1, \cdots, m \end{aligned} \tag{3-10}$$

The matrix formed by $\mathbf{A}$ and $\mathbf{B}$ has to be of rank five, and then if the group of independent rows taken is five, and there are more than five $t_{di}$, these can be found iteratively for the next group of five independent rows of $\tilde{\mathbf{R}}^T\mathbf{S}$. In other words, although [3] performs the algorithm on groups of five, in [2] this is done for $m$ receivers. The choice of five rows at a time is arbitrary, as the result shown in (3-10) holds for $m$ rows.

Once $\mathbf{D}$ is found, then $\tilde{\mathbf{R}}^{\mathbf{T}}\mathbf{S}$ is completely determined and it is possible to find an $\hat{\mathbf{S}}$ and an $\hat{\mathbf{R}}$. These matrices are equivalent to $\mathbf{S}$ and $\mathbf{R}$ up to an affine transformation, a rotation and a translation.

### 3-0-2 Low Rank Approximation

The second algorithm analysed in this research [2], has the purpose of finding the internal delays $t_{di}$. In order to determine these values, the algorithm relies on a low rank approximation of the $\mathbf{R}^T\mathbf{S}$ matrix product.

The problem to be solved can be seen in equation (3-11). The maximum likelihood solution to the problem, is not convex, and therefore will be prone to local minima. Note in (3-11) that the difference $t_{ij} - t_{dj}$ is composed by the measured TOA $t_{ij}$ and the internal delay $t_{dj}$.

$$\sum_{j}^{M}\sum_{i}^{N}(\frac{1}{c}||r_j - s_i|| - (t_{ij} - t_{dj}))^2. \tag{3-11}$$

Then assuming no noise (3-11) leads to the following relationship between the time of arrival and the locations of the receivers and sources ((3-12)).

$$(r_j - s_i)^2 = (c * (t_{ij} - t_{dj}))^2. \tag{3-12}$$

Equation (3-13) is obtained from (3-12) and this is further explained in section 3-0-4 from [4].

$$\mathbf{R^T S} = \mathbf{T} + \mathbf{\Delta}(\delta)\mathbf{W}. \tag{3-13}$$

So a low rank approximation of $\mathbf{R}^T\mathbf{S}$ is to be found, by using a structured total least squares approximation to preserve its structure. A notable difference between the method described in section 3-0-1 and this method, is that this is a recursive method, while [3] has a closed form solution.

### 3-0-3 Autocalibration method

Presented with the problem of finding the onset time estimation this method [7] solves it by using a calibration signal of a known pulse period.

As can be seen in Figure 1-2, the real time of arrival can be determined if the onset times and internal delays are fully known.

By using a calibration signal which is fully determined, then it is possible to find the onset times without error (assuming there is no noise).

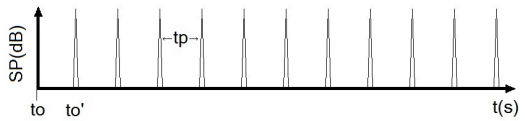



**Figure 3-1:** Calibration wavelet. It depicts a train of clicks emitted at a periodic instances (period $t_p$). The original onset time $t_o$ is unknown, the new onset time coincides with the first received pulse time ($t'_o$)

This can be seen in figure 3-1. The onset time $t_o$ is not known, but the first received pulse time ($t'_o$) is known, as well as the pulse period $t_p$. The measured time of arrival of a signal is $t_{ij} = t_o + TOA_{ij}$, the new $t_o$ is set to be the first received pulse. The following onset time, is equal to the first onset plus one pulse period $t_p$. With this information, then the real $TOA$ can be fully determined.

### 3-0-4 Self calibration method

Once the TOA have been estimated, then the next step is to determine the locations of the receivers and sources. In [4] the developed algorithm outputs the locations of the receivers and sources, with the input of the proper time of arrival. Then the answer comes from a maximum likelihood (ML) problem as in equation 3-14, where $r_i$ represents the position of one receiver, $s_j$ is the position of one source, $TOA_{ij}$ is the time of arrival of a signal from source $j$ to receiver $i$, $c$ is the speed of sound, $n$ is the number of sources and $m$ the number of receivers:

$$\sum_{j}^{n}\sum_{i}^{m}(||r_i - s_j|| - TOA_{ij} * c)^2. \tag{3-14}$$

The minimization of equation (3-14), leads to the estimation of all of the sources and receivers. However, the problem at hand is not convex and may fall in local minima which means the locations would not even be affine with the real locations. The method proposed in [4] aims at eliminating the quadratic terms that arise from (3-14) in an effort to have a bilinear equation on the locations of the receivers and sources which can be solved in closed form.

First, the expanded expression in (3-14) can be seen in (3-15), and there, it is evident that there are quadratic terms for both of the receiver and the source for a given time of arrival:

$$r_i^2 + s_j^2 - 2r_i s_j = (TOA_{ij} * c)^2. \tag{3-15}$$

It is noted that, in order to eliminate the square terms, it is possible to subtract from equation (3-15) the $j = 1$ term for $j = 1, \cdots, n$ *and* $i = 2, \cdots, m$ and then subtract the i=1 term for $j = 2, \cdots, n$ and $i = 2, \cdots, m$. Therefore the resulting equation is bilinear as can be seen in equations (3-16) and (3-17).

$$\begin{aligned}(r_i^2 + s_j^2 - 2r_i s_j) - (r_1^2 + s_j^2 - 2r_1 s_j) - (r_i^2 + s_1^2 - 2r_i s_1) + \\ +(r_1^2 - 2r_1 s_1 + s_1^2) = (TOA_{ij} * c)^2 - (TOA_{1j} * c)^2 - (TOA_{i1} * c)^2 + (TOA_{11} * c)^2\end{aligned} \tag{3-16}$$

And simplifying equation (3-16) gives equation (3-17):

$$\begin{aligned}2r_1 s_i + 2r_j s_1 - 2r_j s_i - 2r_1 s_1 = d_{ij}^2 - d_{1j}^2 - d_{i1}^2 + d_{11}^2 \\ -2(s_i - s_1)(r_j - r_1) = \tilde{t_{ij}}\end{aligned} \tag{3-17}$$

Equation (3-17) can be factored as a product of $(n-1) \times (m-1)$ matrices; $-2\mathbf{R}\mathbf{S}^{\mathbf{T}} = \mathbf{T}$ as can be seen in equation (3-17). In this equation ((3-17)), matrix $\mathbf{R}$ is composed of all 3 dimensional points of the receiver elements $(r_j - r_1)$ and matrix $\mathbf{S}$ contains the positions of all source elements $(s_i - s_1)$.

$$\begin{aligned} \mathbf{R} &\begin{bmatrix} r_{x2} - r_{x1} & r_{y2} - r_{y1} & r_{z2} - r_{z1} \\ & \vdots & \\ r_{xm} - r_{x1} & r_{ym} - r_{y1} & r_{zn} - r_{z1} \end{bmatrix} \\ \mathbf{S} &\begin{bmatrix} s_{x2} - s_{x1} & s_{y2} - s_{y1} & s_{z2} - s_{z1} \\ & \vdots & \\ s_{xn} - s_{x1} & s_{yn} - s_{y1} & sr_{zn} - s_{z1} \end{bmatrix} \end{aligned} \tag{3-18}$$

With the $\mathbf{T}$ matrix fully determined, all $\tilde{d_{ij}}$ are known. The time matrix $\mathbf{T}$ can be decomposed into its singular values, as in equation (3-19).

$$-2\mathbf{R}\mathbf{S}^{\mathbf{T}} = \mathbf{U}\mathbf{V}\mathbf{W} \tag{3-19}$$

Then in order to determine the $\mathbf{R}$ and $\mathbf{S}$ from the decomposition the following expressions holds (4-1):

$$\begin{aligned} \mathbf{R} &= \mathbf{U}\mathbf{C} \\ -2\mathbf{S} &= \mathbf{C}^{-1}\mathbf{V}\mathbf{W} \end{aligned} \tag{3-20}$$

In order to find matrix $\mathbf{C}$, it is assumed that the first receiver and the first source are co-located; $r_1 = s_1$.

### 3-0-5 Noise sensitivity

As a way to compare the localization methods in [2] and [3], both will be analysed in presence of noise. For [3], it has been noted that even small noise levels in the order of $\frac{1}{2f_s}$ $s$ (for $f_s = 48000\ Hz$ ) cause a large error in the internal delay calculations (see Figure 3-2). The effect of noise is also noticeable for the approach followed in [2] as can be seen in Figure 3-3:

If only the fifth receiver is affected by noise, which is not the co-located receiver used in the finding of the positions of receivers and sources, then the resulting errors are:

As it can be seen in Figure 3-3, 3-4 and 3-2, both algorithms have a fair tolerance to noise, with the iterative choice ([2]) performing better. Both of these algorithms are revisited in a following section where instead of basing the calculations on time of arrival, are based on a different the time difference of arrival.

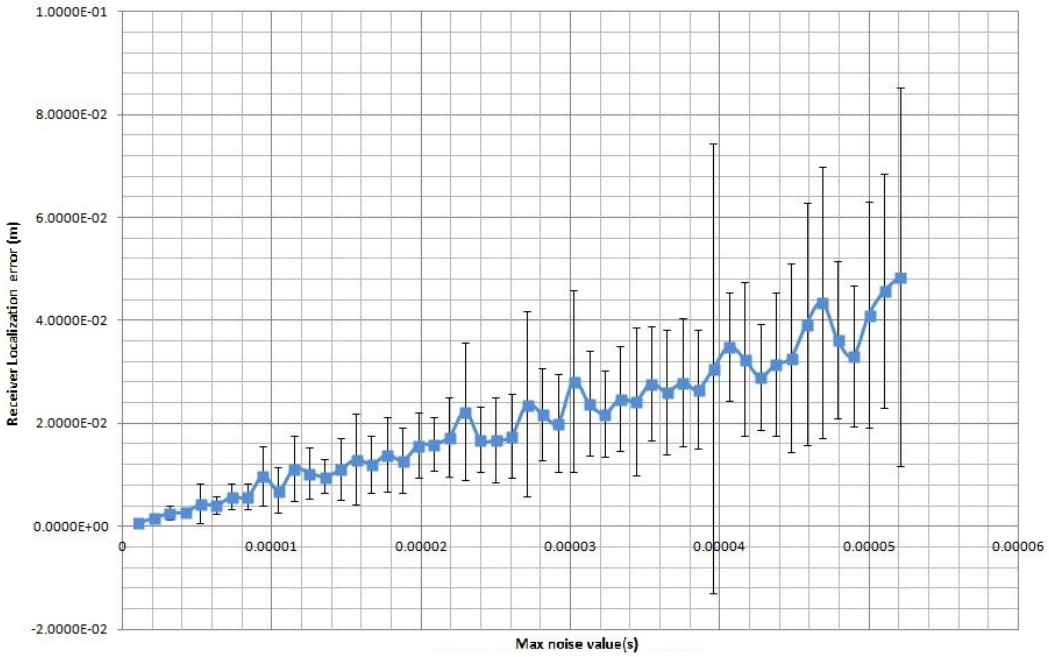



**Figure 3-2:** Location mean error(in meters) caused by noise (using the Pollefeys algorithm).

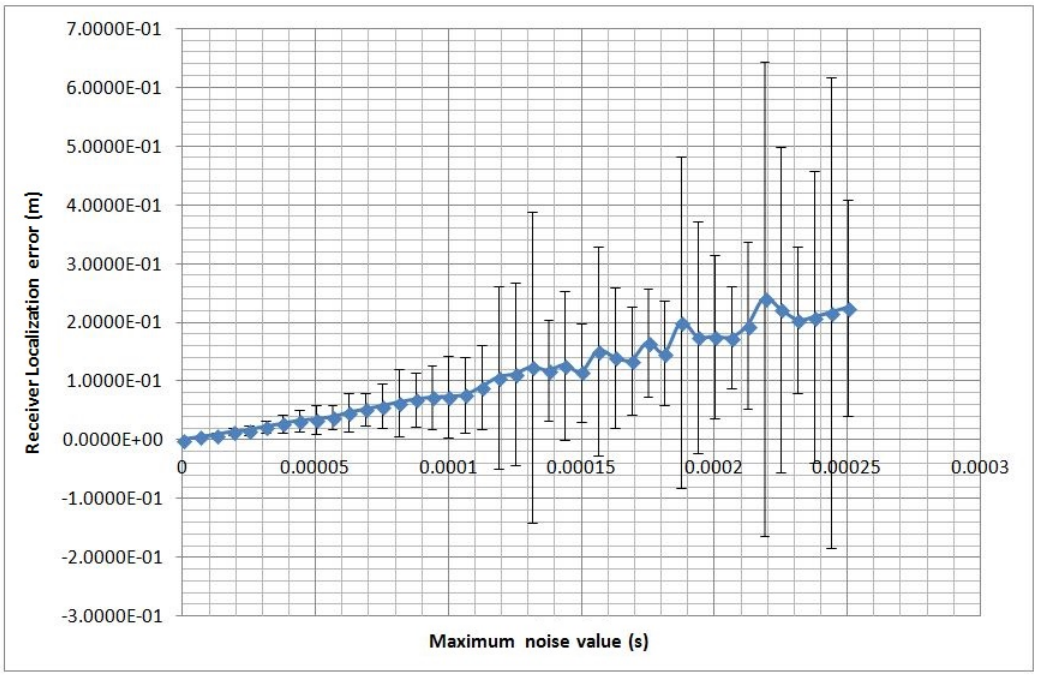



**Figure 3-3:** Location mean error (in meters) caused by noise in the first receiver (using the STLS algorithm).

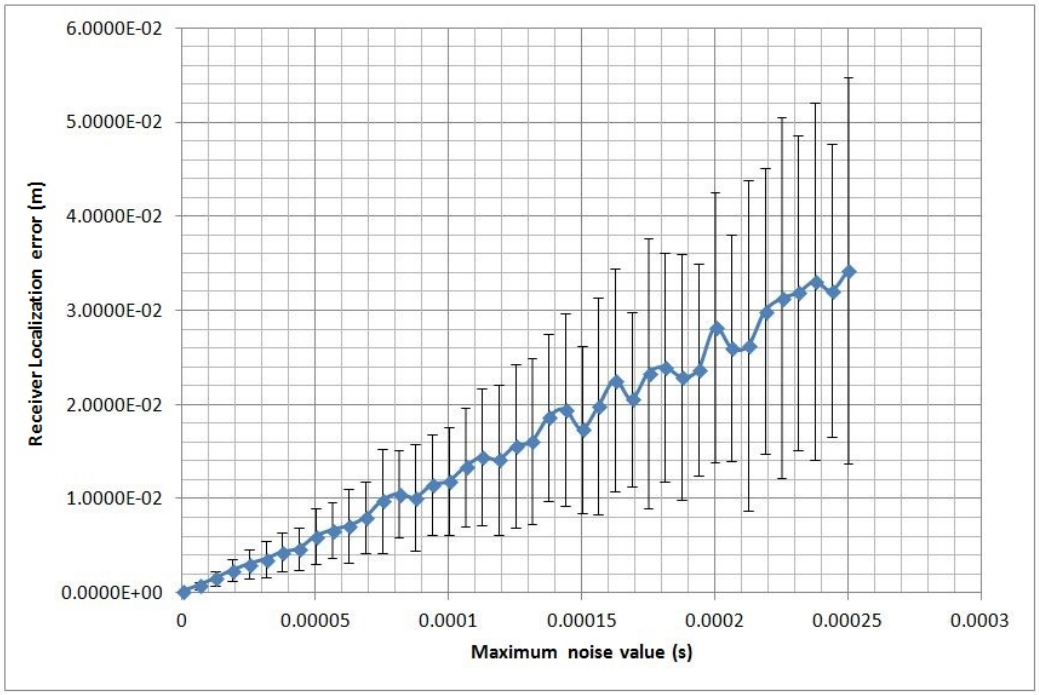



**Figure 3-4:** Location mean error (in meters) caused by noise in the fifth receiver (using the STLS algorithm)

# Chapter 4

# Time Difference of Arrival (TDOA) based approach

In a Section 3, it was explained that when the internal delays $t_{di}$ are unknown, then the localization problem can be solved. This means that the onset times are known $t_{oj}$ or can be found using a fully known calibration signal ([7]).

Then seeing as how the difference between two time of arrival of two different receivers eliminates the onset times (4-1) the goal is to determine how plausible it is to implement these methods using a time difference of arrival (TDOA) approach.

$$\begin{aligned} t_{ij} - t_{kj} &= (t_{oj} + TOA_{ij} + t_{di}) - (t_{oj} + TOA_{kj} + t_{dk}) \\ t_{ij} - t_{kj} &= TOA_{ij} - TOA_(kj) + t_{di} - t_{dk}. \end{aligned} \tag{4-1}$$

As expressed in equation (4-1), the TDOA is equal to the subtraction of the time of arrival of a given signal at two different sensors which can be seen in Figure 4-1.

Assuming that both the internal delays $t_d$ and the onset times $t_o$ are unknown, then it is possible to resort to using the time difference of arrival (TDOA) between the different receivers instead of the times of arrival. Like it was mentioned before the TDOA of a signal between two receivers eliminates the need to calculate the onset time $t_o$. Therefore by reformulating the algorithms ([2] and [3]) from a TDOA standpoint it could could be possible to find the internal delays and with these the location of $r_i$ receivers and $s_j$ sources. It is noted that this difference, though it eliminates the onset times, results in cross terms which is the main challenge when determining if this approach is possible.

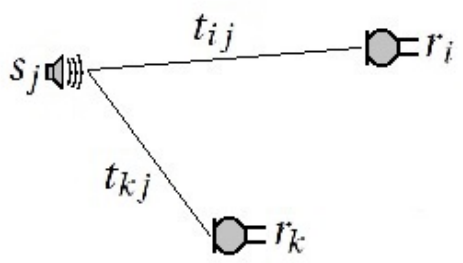



**Figure 4-1:** Time difference between two receivers. The two microphones $r_i$ *and* $r_k$ receive the signal from source $s_j$. The times of arrival at both receivers are $t_{kj}$ and $t_{ij}$

### 4-0-1 Low rank approximation

The first algorithm analysed (3-0-2) relies on a low rank approximation to the time matrix (see (4-2)). In this equation, $\mathbf{R}$ is the matrix containing the positions of the receivers, $\mathbf{S}$ is a matrix which contains the positions of the sources, $\mathbf{T}$ is a matrix which contains the squares of the times of arrival, $\mathbf{\Delta(t_d)}$ is a matrix which contains the individual internal delays and $\mathbf{W}$ contains the times of arrival.

$$\mathbf{R}^T\mathbf{S} = \mathbf{T} + \mathbf{\Delta(t_d)}\mathbf{W}. \tag{4-2}$$

The initial minimization problem, as has been explained in section 1, can be seen in (1-3)

$$\min(\sum^{n}\sum^{R}(||r_i - s_j|| - d_{ij}). \tag{4-3}$$

In the previous expression the distance $d_{ij}$ between receiver $i$ and source $j$ depends on the apparent time of arrival $t_{ij}$, the onset time $t_{oj}$ and the internal delay $t_{di}$ as follows in (4-4):

$$d_{ij} = c(t_{ij} - t_{oj} - t_{di}) \tag{4-4}$$

Then, by using the difference between two TOA from the same source, we can get the TDOA expression:

$$||r_i - s_j|| - ||r_k - s_j|| = v((t_{ij} - t_{di} - t_{oj}) - (t_{kj} - t_{dk} - t_{oj})). \tag{4-5}$$

Then the minimization problem in (4-3) can be reformulated as:

$$\min(\sum_{j}^{n}\sum_{i}^{m}((||r_i - s_j|| - ||r_k - s_j||) - (d_{ij} - d_{kj})). \tag{4-6}$$

And from this expression, we derive the following:

$$(||r_i - s_j||^2 - 2||r_i - s_j||||r_k - s_j|| + ||r_k - s_j||^2) = (t_{ij} - t_{kj})^2 - 2(t_{ij} - t_{kj})(t_{di} - t_{dk}) + (t_{di} - t_{dk})^2. \tag{4-7}$$

The time of arrival problem is solved by exploiting the $\mathbf{R}^T\mathbf{S}$ bilinear product (in the structure of equation (4-2)). This structure should be obtained from the expression $d_{ij} = v(t_{ij} - t_{di} - t_{oj})$, which allows for the separation of positions from the sources and the receivers as in (4-2).

For the approach followed in method [2] the right hand side of (4-2) must be solved for $D(t_{di})$ under the assumption that the internal delays $t_{di}$ are the same during the process for each of the receivers. This can be achieved by means of a low rank approximation. A structured total least squares method is used to find the internal delays for the case where the onset times $t_o$ are known (for example, when using a wavelet generator as sound source 3-0-3). In the case of TDOAs however, the equation is different (see (4-7)). If the equations for $k, i = 1$ and $j = 1$ are subtracted from (4-7) (see [4]), then we arrive at the following expression ((4-8)):

$$\begin{aligned}(||r_i - s_j||^2 - 2||r_i - s_j||||r_k - s_j|| + ||r_k - s_j||^2) - (||r_i - s_1||^2 - 2||r_i - s_1||||r_k - s_1|| + \\ ||r_k - s_1||^2) = (t_{kij}^2 - 2t_{kij}t_{dki} + t_{dki}^2) - (t_{ki1}^2 - 2t_{ki1}t_{dki} + t_{dki}^2).\end{aligned} \tag{4-8}$$

In equation (4-8), the $t_{kij} = t_{kj} - t_{ij}$ represent the difference between two times of arrival $t_{dki} = t_{dk} - t_{di}$, $s_j$ represents the position of one source and $r$ the position of a receiver. The right hand side of (4-8) can be grouped as in (4-9):

$$(t_{kij}^2 - t_{ki1}^2) - (t_{dki})(t_{kij} - t_{ki1}). \tag{4-9}$$

Using the first difference in (4-9) $(t_{kij}^2 - t_{ki1}^2)$ we can form a matrix of size $(m-1) \times n$, the same applies for the $t_{dki}$ and also for $(t_{kij} - t_{ki1})$. This structure is very similar to the structure in (3-13). Then it is necessary to determine if the problem can be solved for $\Delta(t_{dki})$ via a low rank approximation scheme, much like structured total least squares is used in [2]. It is possible to see a similarity between the structure in (3-13) and (**??**) as the squared terms group, and the internal delay differences can be factored from the time differences (see (4-10)).

$$(t_{kij}^2 - t_{ki1}^2) - 2t_{dki}(t_{kij} - t_{ki1}). \tag{4-10}$$

So, stacking the elements in equation (4-10) as in equation (3-13):

$$\mathbf{R}^T\mathbf{S} = \begin{bmatrix} t_{111}^2 - t_{111}^2 & \cdots & t_{11n}^2 - t_{111}^2 \\ \vdots & & \vdots \\ t_{m11}^2 - t_{m11}^2 & \cdots & t_{m1n}^2 - t_{m11}^2 \end{bmatrix} + \begin{bmatrix} t_{11i} & \cdots & 0 \\ 0 & t_{d21} & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & t_{dm1)} \end{bmatrix} + \begin{bmatrix} t_{111} - t_{111} & \cdots & t_{11n} - t_{111} \\ \vdots & & \vdots \\ t_{m11} - t_{m11} & \cdots & t_{m1n} - t_{m11} \end{bmatrix}. \tag{4-11}$$

And now, each element of the left hand side of equation (4-11) can be expressed as:

$$(||r_i - s_j||^2 - 2||r_i - s_j||||r_k - s_j|| + ||r_k - s_j||^2) = T_{q2} - 2\Delta_2(\delta_l)\, W_2. \tag{4-12}$$

It is also possible to factor the left hand side of (4-12) by dividing the expression into two parts. The first contains the squared norms and the other one will contain the cross terms:

$$(\tilde{R}_{ki}^T - 2[\sqrt[2]{(x_k - X_j)^2 + (y_k - Y_j)^2 + (z_k - Z_j)^2}][\sqrt[2]{(x_i - X_j)^2 + (y_i - Y_j)^2 + (z_i - Z_j)^2}]) - \tilde{S}_j, \tag{4-13}$$

where:

$$\tilde{R}_{ki} = \begin{bmatrix}[r_k^2 + r_i^2] & -2(x_k + x_i) & -2(y_k + y_i) & -2(z_k + z_i) & 2\end{bmatrix}^T$$
$$\tilde{S}_j = \begin{bmatrix}(1 \cdots 1) & X_j & Y_j & Z_j & s_j^T s_j\end{bmatrix}^T.$$

Then equation (4-12) can be rewritten as:

$$\hat{r}_{ki}^T\, \tilde{s}_j = T_{q2} - 2\Delta_2(\delta_l)\, W_2, \tag{4-14}$$

where:

$$\hat{r}_{ki} = \begin{bmatrix}[r_k^2 + r_i^2] - 2||r_i - s_j||||r_k - s_j||] & -2(x_k + x_i) & -2(y_k + y_i) & -2(z_k + z_i) & 1\end{bmatrix}^T.$$

These two terms can be stacked into matrices $\hat{\mathbf{R}}$ and $\tilde{\mathbf{S}}$. In order for the algorithm to work, both matrices have to be of known rank $r$ and $r \leq d$, which means that $r$ represents the dimension of the space where the array of sensors is located. The first one is a $m\text{x}(n-1)+5$ matrix, while the second is $(n-1)+5\text{x}m$. So, the product $\hat{R}^T\, \tilde{S}$ has a rank 5, and the product is not bilinear, as $\hat{R}^T$ has source position elements ($s$) in it. Thus, because of the cross terms which arise from the time difference of arrival, the structured total least squares formulation is not adaptable to the TDOA framework.

### 4-0-2 Pollefeys's method

As was explained in section 3-0-1, the method proposed by [3] takes advantage of the structure in the left hand side of (3-8). Being able to use the product $\tilde{\mathbf{R}}^{\mathbf{T}}\mathbf{S}$ depends greatly on the fact that the time of arrival (TOA) of the signals is known, which is not a likely case. However, it is possible to rearrange (4-7) as follows:

$$(||r_i - s_j||^2 + ||r_k - s_j||^2) = (t_{ij} - t_{kj})^2 - 2(t_{ij} - t_{kj})(t_{di} - t_{dk}) + (t_{di} - t_{dk})^2 + 2||r_i - s_j||||r_k - s_j||.$$

And this is equal to:

$$\tilde{R}_{ki}^T \tilde{S}_j = (t_{ij} - t_{kj})^2 - 2(t_{ij} - t_{kj})(t_{di} - t_{dk}) + 2||r_i - s_j||||r_k - s_j||, \tag{4-15}$$

where:

$$\begin{aligned} \bar{r}_{ki} &= \begin{bmatrix} [r_k^2 + r_i^2] - (\delta_k - \delta_i)^2 & -2(x_k + x_i) & -2(y_k + y_i) & -2(z_k + z_i) & 2 \end{bmatrix}^T \\ \bar{s}_j &= \begin{bmatrix} 1 & X_j & Y_j & Z_j & s_j^T s_j \end{bmatrix}^T. \end{aligned}$$

With $t_{dki} = (t_{dk} - t_{di})$ and $d_{ij} = ||r_i - s_j||$, the right hand side of (4-15) can be factorized as follows:

$$T_{ikj}^2 + \begin{bmatrix} t_{dki} & d_{ij} d_{kj} | \end{bmatrix} \begin{bmatrix} -2T_{kij} \\ 2\, I \end{bmatrix} = T_{ikj}^2 + \tilde{\Delta} T.$$

Then, after stacking up (4-15), this closely resembles the structure seen in (3-8):

$$\tilde{\mathbf{R}}^{\mathbf{T}} \tilde{\mathbf{S}} = \mathbf{T2} + \mathbf{\Delta}(\mathbf{t_d}, \mathbf{d})\mathbf{T} \tag{4-16}$$

Where:

$$\mathbf{\Delta} = \begin{bmatrix} t_{d11} & \cdots & 0 & d_{11}d_{11} & d_{12}d_{12} & \cdots & d_{1n}d_{1n} \\ \vdots & \ddots & \vdots & d_{11}d_{21} & d_{12}d_{22} & \cdots & d_{1n}d_{2n} \\ 0 & \cdots & t_{d1m} & d_{11}d_{m1} & d_{12}d_{m2} & \cdots & d_{1n}d_{mn} \end{bmatrix},$$

$$\mathbf{T2} = \begin{bmatrix} t_{111}^2 & t_{112}^2 & \cdots & t_{11n}^2 \\ t_{121}^2 & t_{122}^2 & \cdots & t_{12n}^2 \\ \vdots & \vdots & \cdots & \vdots \\ t_{1m1}^2 & t_{1m2}^2 & \cdots & t_{1mn}^2 \end{bmatrix} \textit{and,}$$

$$\mathbf{T} = 2\begin{bmatrix} -t_{111} & -t_{112} & \cdots & -t_{11n} \\ -t_{121} & -t_{122} & \cdots & -t_{12n} \\ \vdots & \vdots & \cdots & \vdots \\ -t_{1m1} & -t_{1m2} & \cdots & -t_{1mn} \\ 1 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Then as in [3], $\tilde{\mathbf{S}}$ generates the row space of $\mathbf{T2}+\boldsymbol{\Delta}\ \mathbf{T}$, which means that a linear combination of the right hand side of equation (4-16) can result in any of the rows of $\tilde{\mathbf{S}}$. Since the last row of $\tilde{\mathbf{S}}$ is a row of ones, then the following holds:

$$C[\mathbf{I} \quad \boldsymbol{\Delta}] \begin{bmatrix} \mathbf{T2} \\ \mathbf{W} \end{bmatrix} = [\mathbf{1} \ \cdots \ \mathbf{1}] \tag{4-17}$$

And then, finding $[C\mathbf{I}\ \mathbf{C}\boldsymbol{\Delta}]$ can be achieved by taking the pseudoinverse of $[\mathbf{T}, \mathbf{W}]^\top$:

$$C[\mathbf{I} \quad \boldsymbol{\Delta}] = [\mathbf{1} \ \cdots \ \mathbf{1}] \begin{bmatrix} \mathbf{T2} \\ \mathbf{W} \end{bmatrix}^+$$

Once $X = [cI\ c\Delta]$ is known, and with the knowledge of the size of $\mathbf{I}$ similarly as in section 3-0-1, each unknown is found as a piecewise division.

Since the first $m$ columns of $\boldsymbol{\Delta}$ contain the $t_{dki}$ in a diagonal, and because $cI$ is a vector of $m$ elements, then by dividing each $c_i\ \delta_l$ from $c\Delta(\delta)$ by each element of $cI$ as:

$$\boldsymbol{\Delta}(\mathbf{k}) = \frac{\mathbf{X_{k+5}}}{\mathbf{X_k}} \tag{4-18}$$

The results of 4-18 however, are not unique and then the values of $t_{dk}$ are not the values of the internal delays and then the estimation of the locations of receivers and sources fails. The reason why all $t_{dk}$ are not a unique solution to 4-18 is that the rank of $\left[\mathbf{A}//\mathbf{B}\right]$ is greater than $m$ (the number of sensors). In order to have a unique solution, the rank of the mentioned matrix $\boldsymbol{\Delta}$ has to be, at most $m$. From equation (4-17), it is possible to see that, in order to find a unique $\boldsymbol{\Delta}$, the second element has to be a tall matrix, but in reality both T and W are fat matrices ($nmx2n$) due to the inclusion of the cross-term unknowns. Then, the $\boldsymbol{\Delta}$ found is not unique.

# Chapter 5

# Arbitrary sound sources

Since the use of time difference of arrivals in current algorithms did not yield the desired results, a new approach to estimating the onset times is proposed. The purpose of this approach will be again, to find the onset times $t_o$. Once the onset times are found then it is possible to estimate the locations of receivers and sources using known methods $t_d$ ([3], [2]).

In [7] the authors devise a way to find the onset time which employs a calibration source of known frequency as is explained in section 3-0-3. There is a limitation in this approach which derives from the fact that the calibration signal has to be known beforehand, and this implies that a device which generates this signal has to be moved around the receivers (for example, a wavelet generator).

$$t_{oj} + \frac{||r_i - s_j||}{c} + t_{di} = t_{ij}, \tag{5-1}$$

where $t_{oj}$ is the onset time at source $j$, $c$ is the speed of sound and $t_{di}$ the internal delay at receiver $i$. So, from (**??**) it is clear that if the problem of finding $r_i$ and $s_j$ is to be solved using times of arrival, then the $t_{oj}$ has to be known.

A way around the limitation imposed by the need to use such a wavelet generator, is to use a periodic or quasi-periodic signal as a calibration source. It so happens that voiced speech is a quasi-periodic signal ([8]), and the periods of the excitation signal of voiced speech can be estimated very accurately ([1], [9]).

Then, by taking voiced speech as a calibration signal, and estimating its pitch periods, it is possible to estimate the onset times of each of these periods in order to get an estimate of the onsets $t_{oj}$.

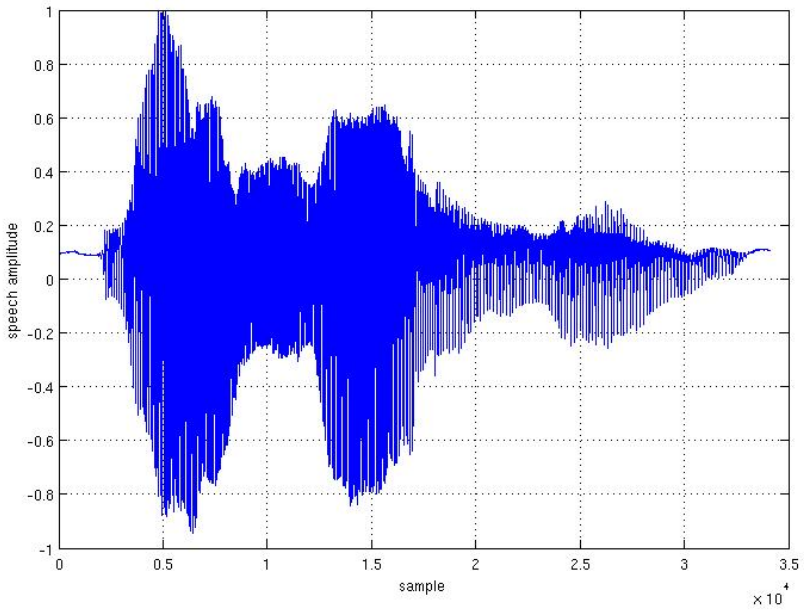



**Figure 5-1:** Example of a voiced-unvoiced speech segment, sampled at 48 kHz.

## 5-1 Localization using voiced speech signals

In section 4 it was clear that using time differences of arrival does not solve the localization problem when both the onset time ($t_o$) and internal delays ($t_d$) are unknown. Therefore finding the onset times $t_o$ is done in [7] with the use of a calibration signal. This signal is known a-priori and using the knowledge of its $t_p$ period it is possible to find the onset times with good accuracy. Since voiced speech signals can be modelled as a filtered impulse train of a given frequency, then those signals are good candidates to be used as a calibration signal.

$$t_{ij} = \frac{||r_i - s_j||}{c} + t_{oj} + t_{di}, \tag{5-2}$$

where $t_{oj}$ is the onset time at source j and $t_{di}$ is the internal delay at receiver i.

As suggested by [7], the $t_j$ terms can be estimated if the calibration signals are known, that is, the periods $t_p$ at which that signal is emitted are known. Then, for every successive $j$, the onset time is $t_{o1} + jt_p$:

$$t_{ij} = \frac{||r_i - s_j||}{c} + (t_{o1} + jt_p) + t_{di}.$$

Using speech signals for this purpose, has an important implication; the pitch is not a constant. Therefore $t_p$ is not a unique value and if the calibration signal is completely known, then each of the intervals $t_{pj}$ is also known. With this in mind (5-2) can be rewritten as:

$$t_{ij} = \frac{||r_i - s_j||}{c} + (t_1 + \sum_{k=2}^{j} t_{pk}) + t_{di}. \tag{5-3}$$

In this scenario, a single person produces voiced speech and moves relative to the receivers at a rate much lower than $t_p$. Once the signal has been received it is necessary to find the period of the signal. Pitch period is estimated in [1] by finding the glottal closure instances (GCI) of the speech signal and the difference between two of such GCI gives the pitch period at that instance.

Then the onset time at a given source location is equal to the sum of all previous onset times as stated in equation (5-3). There is still an unknown, namely the initial onset time $\tau_0$. However, if this initial onset is assumed to be zero, then equation (5-3) is modified as:

$$t_{ij} = \frac{||r_i - s_j||}{c} + (\sum_{k=2}^{j} t_{pk}) + t_{di}. \tag{5-4}$$

The remaining source of error in this case, is related to sampling. If the information about the calibration signal is known completely (the type of signal and its $t_p$) then the estimation of the onset times only has that problem.

When using a voiced speech signal this assumption is not true, as the $t_p$ will also have to be determined from the received signal. A problem with this is that the time at which each pulse is detected also includes the actual time of arrival and the $t_{di}$ which is also unknown. It also changes depending on the $i-th$ receiver and depending on the position of the source if it is moving. Therefore the difference between peak times is not $t_p$. There must be a way to retrieve $t_p$ from the received data ((5-4), (5-5)) under a set of well defined conditions,

$$\hat{t_{pj}} = \tilde{t_{ij+1}} - \tilde{t_{ij}} = (t_{ij+1} + (j+1)t_p) - (t_{ij} + jt_p). \tag{5-5}$$

Then, if the receiver positions and the source positions are assumed to be normal distributed then the following holds:

$$\frac{1}{M}\sum_{i=1}^{m} \hat{tp_j} = \frac{1}{m}\sum_{i=1}^{m}[(t_{ij+1} + (j+1)t_p) - (t_{ij} + jt_p)] = \sum_{i=1}^{M}(t_{ij+1} - t_{ij}) + t_p. \tag{5-6}$$

So, the location of the receivers are assumed to be zero mean. This implies that some of the receivers appear to be closer (when the source approaches those), and other sources appear to be further away (when the source moves away). This can be seen in Figure 5-2, where each color represents a single glottal closure instance and each type of point (star, square, x, etc.) represents an individual receiver. As time progresses, since the source is moving, these points shift relative to each other.

Then, the mean of the measured times of arrival tends to zero as the number of receivers increases $\sum_{i=1}^{m}(t_{ij+1} - t_{ij}) \approx 0$ and then (5-6) can be rewritten as (with an error $\epsilon$):

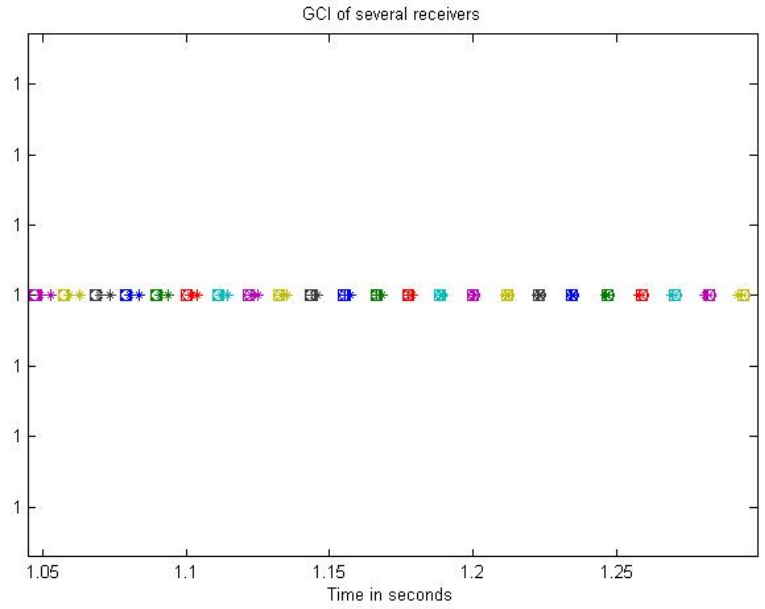



**Figure 5-2:** Voiced-unvoiced speech sample. Each different point (star, square, circle, etc) represents the time of arrival of a single pulse at a different receiver

$$\sum_{i=1}^{m}(t_{ij+1} - t_{ij}) + t_p = t_p + \epsilon. \tag{5-7}$$

Because of these conditions, there are several considerations to be taken into account, regarding the locations of receivers and sources, as those considerations have a direct impact on the results of the estimation of both receivers and sources.

It is convenient to describe the ideal setup of the system. First, there are $m$ sensors and $n$ sources, which are distributed in an ideal room (with no reverberations). Second, the $n$ sources are actually a single source, which moves around in the room. This trajectory is also restricted; as it has been deemed illogical to assume that the source moves across the sensors or does so in a purely random way.

Therefore, for this project, the path followed by the sources is modelled as a curve and also as a helix (Figure 5-3). The latter one, although unrealistic, works as a proof of concept. The sensors remain in the center of the trajectory placed using a zero mean distribution.

An alternative setup was also tested, and it is aimed at a scenario in which several sources are quasi-static; that is, they only move in a semicircular path with a radius of 0.5 meters. This is aimed at modelling what it would be like if the speech signals from several people where used as calibration signals, without each subject moving around the receivers like in the previous scenario.

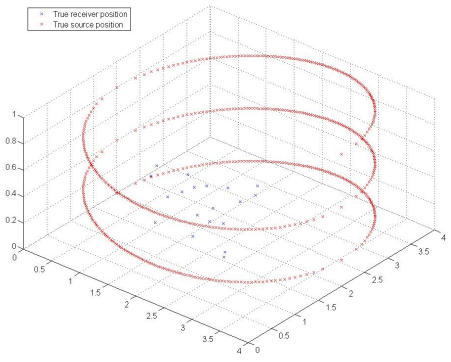



**Figure 5-3:** Location of sensors in blue and the instances of the moving source in red

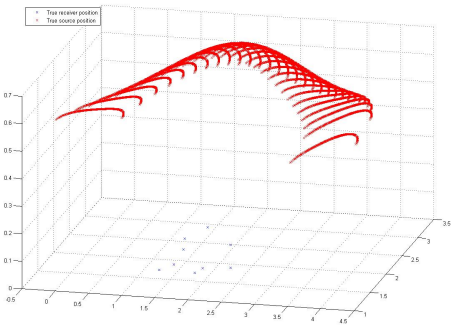



**Figure 5-4:** Location of sensors in blue and the instances of the quasi-static sources in red

### 5-1-1 Test calibration signal with fixed period

Considering the distribution mentioned in section 5-1 of the sources and the receivers the first approach to the voiced signal is to assume that the sources are emitting a pulse train with a fixed frequency; this means that $t_p$ is a constant:

$$t_{ij} = \frac{||r_i - s_j||}{c} + (t_{p1} + jt_p).$$

With these constraints, and assuming that the frequency of the calibration signal is known, then the time of arrivals can be found. This assumption is certainly the trivial case, and then the first task is to try to determine the $t_p$ from the received signal.

Again, going to (5-5) all microphones will receive the same pulse at a slightly different time, and the distance between pulses at a given receiver changes depending on the movement of the source. However, because of the assumption that the receivers follow a zero mean distribution, as per (5-6), the value of $t_p$ can be found from two consecutive measurements with reasonable accuracy. Then, the estimation of the $t_p$ gives an estimation of $t_{ij}$ which is then used in [4] to find the locations of both receivers and sources.

### 5-1-2 Speech calibration signals

Assuming that the calibration signal is a pure pulse train, even if the frequency of the pulses varies in time is impractical. Such pulses imply that a generator must be used and in such cases the signal would be completely known. If the case was to use a synthetic calibration signal, then it would not be necessary to estimate the pulse period and the signal would be completely characterized as in [7]. Then the main objective is to be able to use a readily available calibration source such as speech. Once more, the assumption is made that only one source is present and it moves according to an unknown trajectory, which is not random (for example in a circle around the receivers).

Speech is a quasi-periodic signal, and a long speech sequence is normally divided in sections of voiced and unvoiced speech. Voiced speech has a quasi-periodic excitation signal, while unvoiced speech is excited by gaussian noise (see figure 5-5).

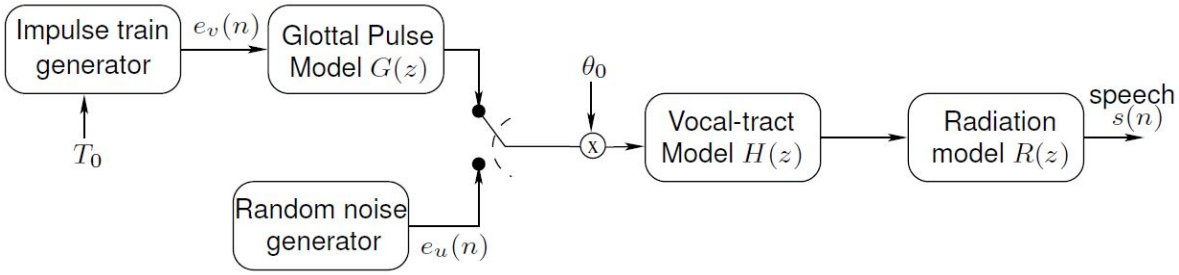



**Figure 5-5:** Block model of voiced-unvoiced speech signals

Therefore, it is necessary to determine the instantaneous pitch frequency from the received signal in order to estimate the actual $t_{pj}$. In order to find the pitch period an algorithm named GEFBA is used ([1]). This algorithm combines a speech presence detector with a glottal flow derivative (GFD) approach to estimate the GCI (Glottal closure instances) of the voiced segments.

The first approach is to assume that the segment received is composed only of voiced speech. Then the all of the received sequence will be useful and there is no need to separate the signal between voiced (from which GCIs are extractable) and unvoiced speech. This step makes the unrealistic assumption that a received speech signal can contain only voiced speech and was used as a intermediate step towards the use of a full voiced-unvoiced speech segment.

In the most realistic approach, it is assumed that a speech sequence is received, and this sequence contains both voiced and unvoiced speech segments. In this case it is first necessary to determine whether or not the sample is part of the voiced or unvoiced part of the speech signal. Then, from the voiced speech parts, it is necessary to extract the GCIs and from these detected time instances, it is possible to estimate the onset times.

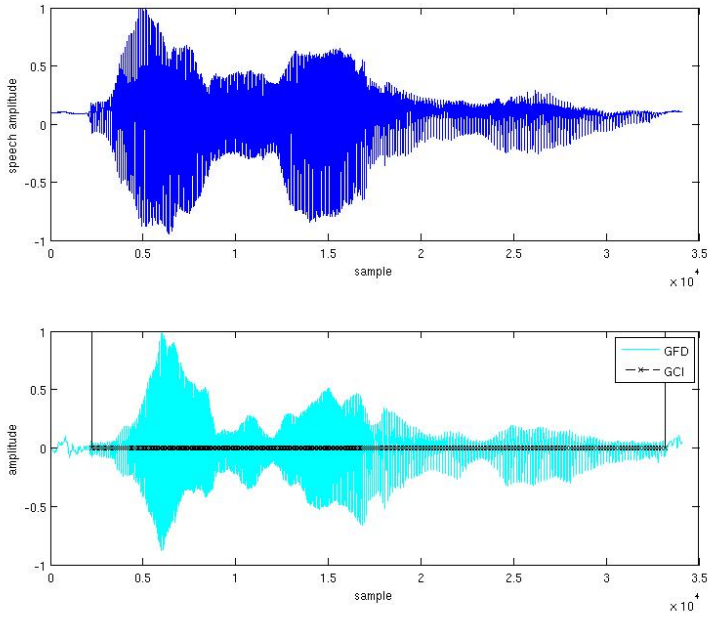



**Figure 5-6:** Plot of a purely voiced speech signal [Top]. Glottal flow derivative with Glottal Closure Instances (GCI) [Bottom]

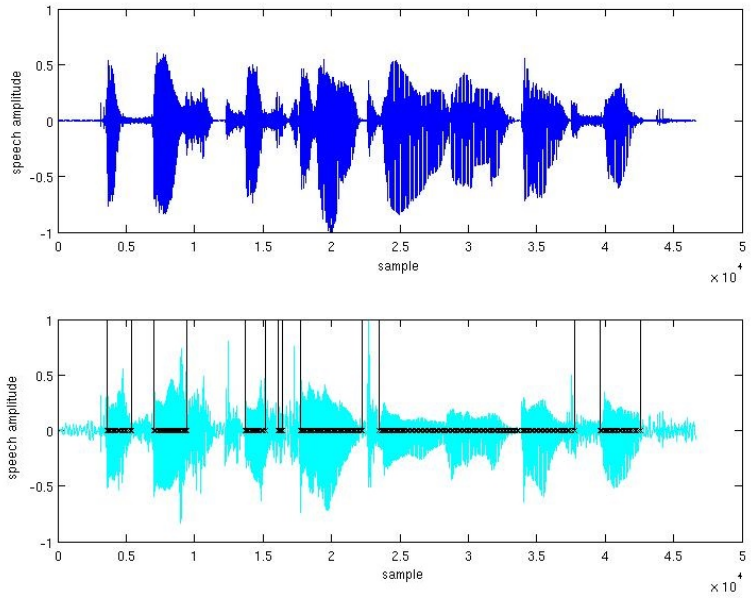



**Figure 5-7:** Plot of a voiced-unvoiced speech signal [Top]. Glottal flow derivative with Glottal Closure Instances (GCI) [Bottom]

# Chapter 6

# Simulations and Results

For the purpose of simulation, several assumptions have to be taken into account. It was already mentioned in section 5-1 that a crucial assumption is that the room impulse response is null; which means that the reverberations are equal to zero. Then the simulation consists of a voice sequence being emitted by a moving source and the source will be at a given location during the onset of every GCI. Then taking into account the precise distance of the source at the instance of the onset of that CGI it is possible to know the time necessary for sound to cover the distance between that source location and each of the receivers. With that information, it is then possible to properly determine the TOA of the signal, and this calculation is performed for each pulse of the speech signal. This is equivalent to a fractional delay calculation of the moving source as is done in [10] and [11], but instead of the signal, what is modelled is the arrival times of each CGI of the speech signal. These times include the pitch period and the TOA. Another way to look at this, is to assume that each GCI is an independent source which only emits one pulse.

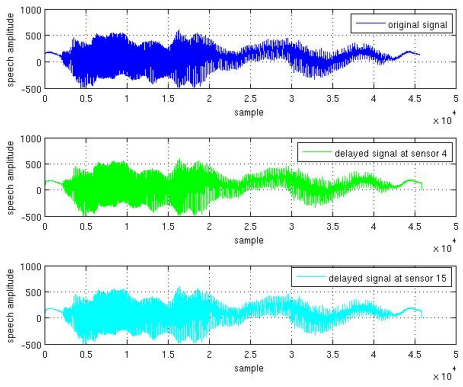



**Figure 6-1:** Voiced samples at the source [top], receiver 4 [middle] and receiver 15 [bottom]

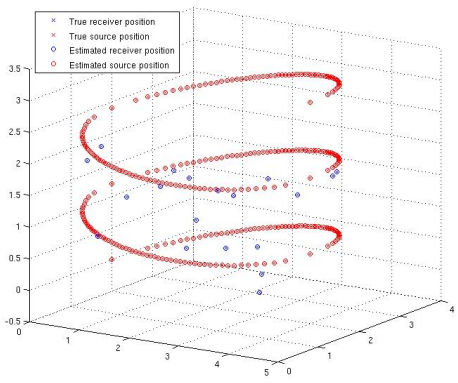



**Figure 6-2:** Calculation of source (red) and receiver (blue) locations when offsets are known.

$$T_{ij} = T_(pj) + TOA_{ij} + \epsilon_Q. \tag{6-1}$$

Initially it is assumed that there is no quantization error $\epsilon_Q$ but even when it is taken into account because it is very small (in the order of $\frac{1}{2F_s}$), it adds very little to the final error.

### Pulse train

As mentioned in section 5-1-1, the first approach to the problem was done using a test pulse train. First the pulse train had a fixed frequency, and then the frequency was time dependent (a chirp). The simulation assumed a random distribution of the receivers, with the provision that the distribution has a mean equal to zero. When the period of the pulse train is known, then the locations are perfectly determined, regardless of where the sensors are located (see Figure 6-2)

When the $T_a$ is unknown and it has to be estimated (5-6), then the resulting estimation of the locations is not perfect. The error in the estimation of the offsets (and therefore of the locations) is influenced severely by changes in the location of the receivers; i.e. if the receivers are not inside of the path drawn by the moving source, then the error quickly escalates.

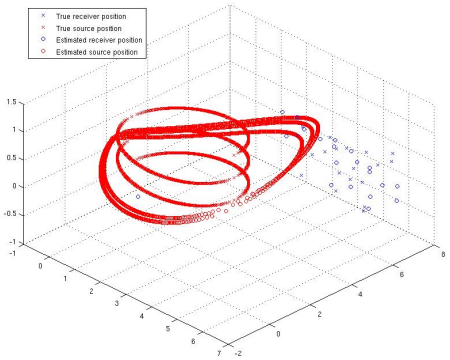



**Figure 6-3:** Calculation of source and receiver locations when offsets are unknown and the receivers are outside of the path of the moving source.

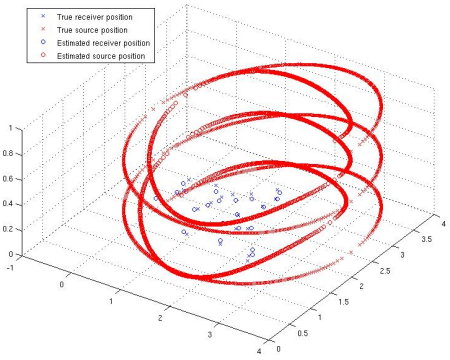



**Figure 6-4:** Calculation of source and receiver locations when offsets are unknown and the receivers are in the space enclosed by the path of the moving source.

If the sources are simulated to be inside of the closed path of the receiver, then the estimation is much more accurate.

### Speech signals

The motivation for the use of speech as calibration signal is mostly the convenience and availability it has. While the use of a wavelet generator of known pulse will completely solve the onset times, the use of such a device can be problematic especially when such a device

is not available at a given moment. Similarly to the case of a pulse train if the pitch period is known the location of the sources and receivers is completely determined. But, since this assumption is not realistic, the estimation of such periods also depends on the arrangement of the receivers. If the the locations of the sensors is zero mean distributed, then the estimation can be performed with a good degree of accuracy as can be seen in the estimation of the locations using a speech segment 6-3.

In an even more realistic setting, the source is not likely to move in a helical trajectory, but it more probably will follow a semicircular path around the receivers. In such a case, the error increases if compared to the helical movement of the source but it still remains reasonably low provided that the sensors are effectively arranged with a mean equal to zero and surrounded by the path drawn by the moving source, as can be seen in Figure 6-6.

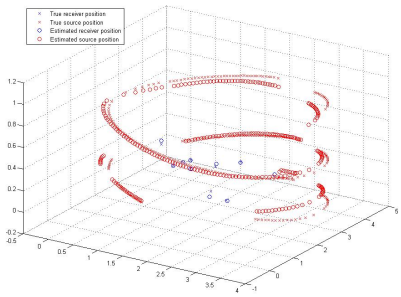



**Figure 6-5:** Calculation of source and receiver locations when offsets are unknown and the receivers are in the space enclosed by the helical path of the moving source.

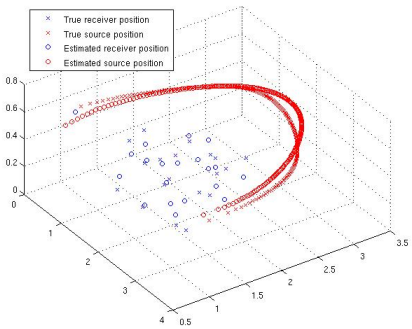



**Figure 6-6:** Calculation of source and receiver locations when offsets are unknown and the receivers are in the space enclosed by the semicircular path of the moving source.

The effects of noise in the measurements, which we assume is due to quantization error (which is a function of the sampling frequency $\epsilon_s = \frac{2}{fs}$), has an effect on the results of the localization of the sources and receivers. There is a difference in the resulting errors when the calibration signal is only voiced speech or if it has voiced and unvoiced parts and the path followed by the source also has an effect.

Also, if the source moves around the receivers in a closed circle the error is kept low, close to the error achieved by the helical trajectory of the moving source.

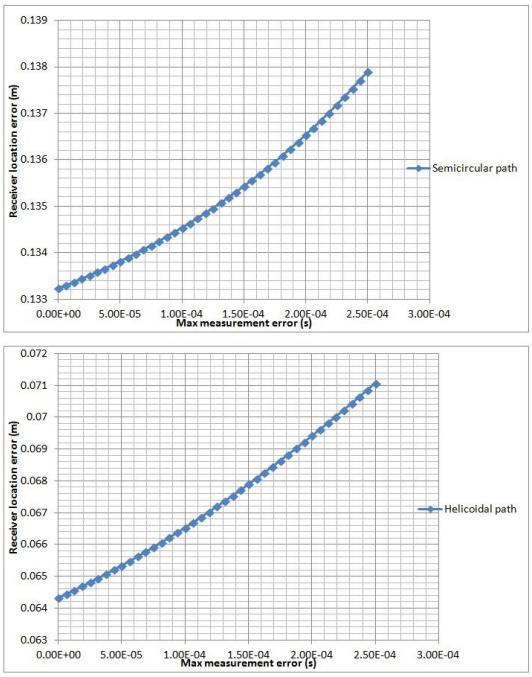



**Figure 6-7:** Localization errors caused by quantization error with voiced only speech for a semicircular path (top) and helicoidal path (bottom) .

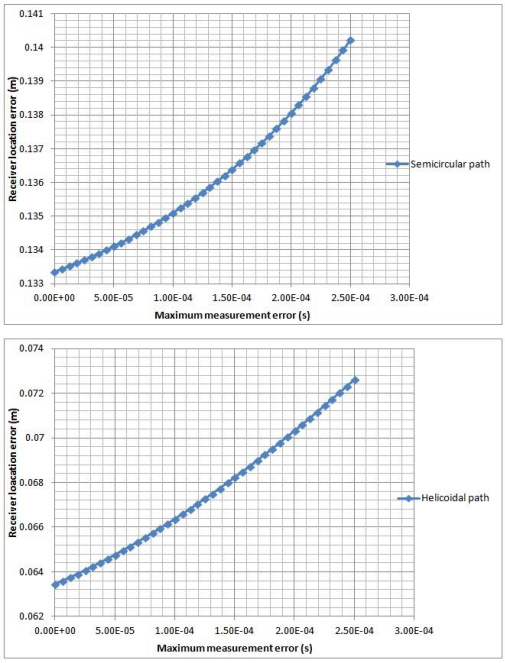



**Figure 6-8:** Localization errors caused by quantization error with voiced and unvoiced speech for a semicircular path (top) and helicoidal path (bottom).

Another setup that tested assumes a number of quasi-static sources, which move in a small radius and are distributed in a semicircular path around the receivers, can be seen in Figure (5-4). In this case, each of the sources which moves about with a radius of 0.5 meters is used to calculate the position of the receivers. The individual errors of each of these is significantly large (see Figure 6-9 and Figure 6-10):

In Figure 6-10, all the estimated locations using the individual sources which produce speech signals are shown. All sources are relatively close to each other and they move identically in a semicircle therefore facing always in the same direction. This also brings forward another difference in errors; if the semicircle faces the receivers as in Figure 6-12 the errors are significantly lower than the errors obtained when the sources face elsewhere (see Figure 6-9:

This also leads to the final setup in which individual sources move randomly in a small area and these individual sources are set along a circle or semicircle around the receivers. This aims at reproducing a number of persons who move their head randomly in an area of known size (using a gaussian distribution) while they produce speech, with every source producing speech uninterrupted by other sources. This means that each person would speak alone, until every source has produced the speech needed for the localization process to take place. This distribution can be seen in Figure 6-13 and Figure 6-14

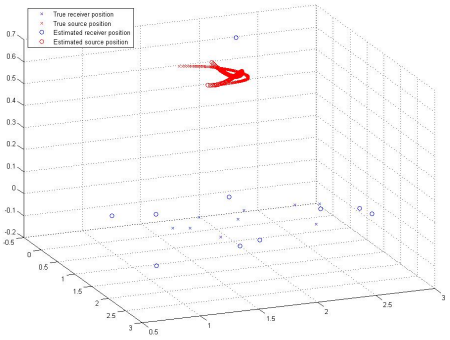



**Figure 6-9:** Individual localization error caused by quantization error with voiced and unvoiced speech for one quasi-static source only.

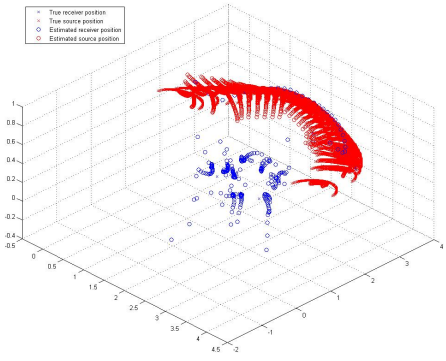



**Figure 6-10:** Individual localization error with voiced and unvoiced speech for one quasi-static source only.

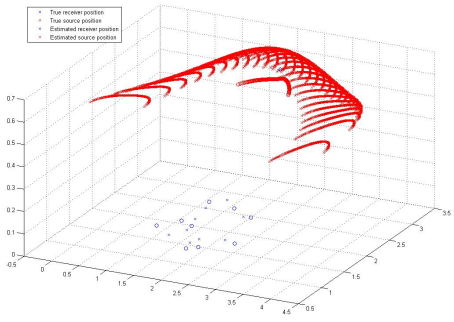



**Figure 6-11:** Localization errors with voiced and unvoiced speech.

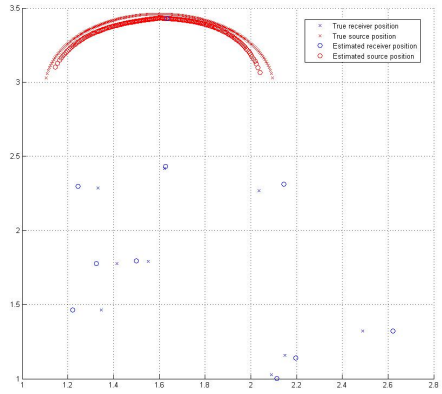



**Figure 6-12:** Individual localization error with voiced and unvoiced speech for one quasi-static source only.

## 6-0-1 Comparative results

### Effect of parameter adjustments

It is expected that, apart from the relative positions and distribution of sources and receivers, the number of sensors and sources, has an effect in the results. The number of sources is large when speech is used, as the typical fundamental frequency of a male speaker ranges from 80 Hz to 170 Hz. This means that in a single second there will be somewhere between 80 and 170 individual sources available. Therefore the main concern is how the results are affected by the number of receivers.

From Figure 6-15, Figure 6-16 and Figure 6-17, it is clear that as the number of receivers increases then the error decreases. After a number of receivers the decrease in error is quite small. This is consistent with the assumptions made regarding the distribution of the receivers.

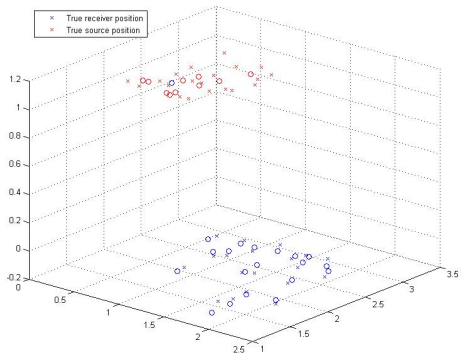



**Figure 6-13:** Individual localization error with voiced and unvoiced speech for one quasi-static random source only

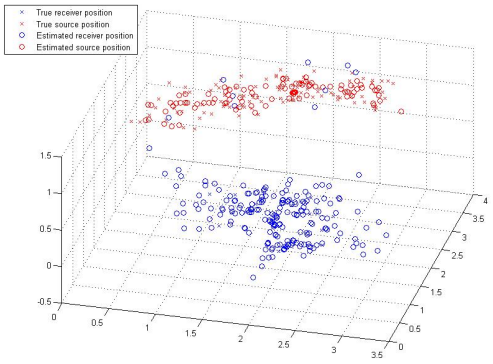



**Figure 6-14:** Localization error with voiced and unvoiced speech with all quasi-static random sources shown at once

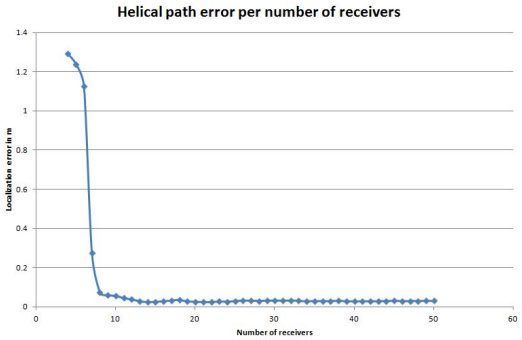



**Figure 6-15:** Localization error based on the number of receivers for a helical source path

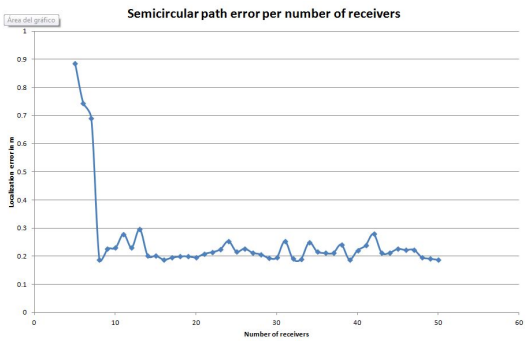



**Figure 6-16:** Localization error based on the number of receivers for a semi-circular source path

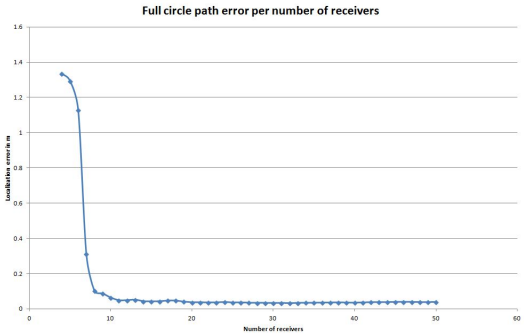



**Figure 6-17:** Localization error based on the number of receivers for a circular source path

### 6-0-2 Implementation

The implementation was performed with an array of microphones and a person speaking while moving in a given trajectory (e.g. see figure 6-22). The array consisted of up to eight microphones arranged in a table of 70 cm, distributed in a space of 60 cm by 55 cm by 5.5 cm. The microphones are quite directional, so they were positioned upwards in an effort to make their pick-up as isotropic as possible. Initially measurements were carried out in an uninsulated room close to air conditioning systems (see fig 6-18 and fig 6-19). As a result there was a particularly high noise level and high reverberation from the walls of the room, and this causes a high discrepancy in the number of glottal closure instances that were detected and thus affected the final results.

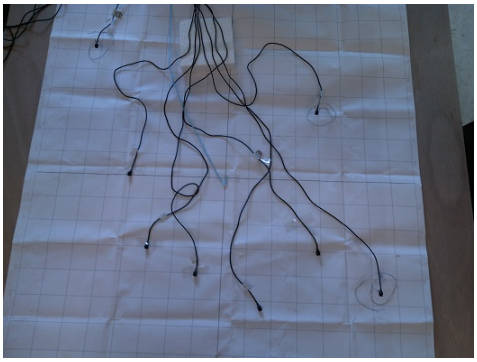


**Figure 6-18:** Initial setup in uninsulated room

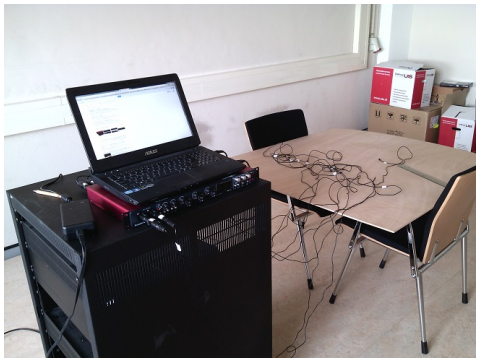


**Figure 6-19:** Initial receiver distribution

The fix to this problem was to carry out the experiments in an environment which greatly minimizes the impact of reverberation and ambient noise (ventilation systems, computer fans, motor vibrations, etc.), which in this case is an anechoic chamber. The setup used in this case can be seen in Figure 6-21 and Figure 6-22, where the microphones and the sound absorbing walls and floor are visible. The path followed was a half circle and full circle of 1.5 m of radius (see Figure 6-20).

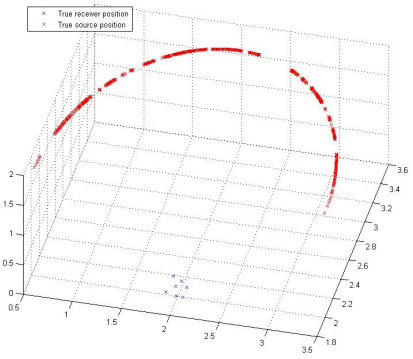



**Figure 6-20:** Experimental setup

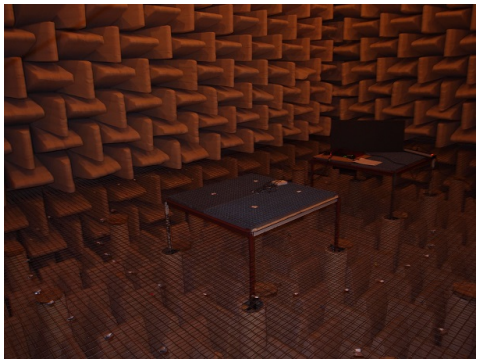


**Figure 6-21:** Experimental setup in anechoic chamber

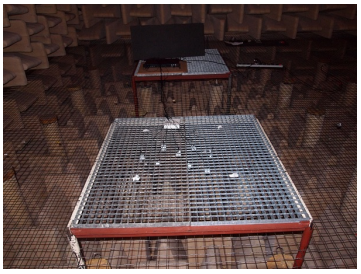


**Figure 6-22:** Receiver distribution

The speech is recorded at the 8 microphones (AKG C417 lapel microphones) and the sound is picked by a Focusrite Scarlett 18i20 USB audio interface. The data sampled at 48 kHz and is saved in a matrix, with each row in this matrix being the recording of moving speech at each of the microphones. The resulting estimations can be seen in Figure 6-23 and in Figure 6-24.

The average error of the experiments is of $39cm$, however, it can be seen in Figure 6-23 and in Figure 6-24 that the structure of the receivers is preserved up to a rotation, a scaling and a translation. It was determined, after the experiments had been carried out, that one condition of the algorithm had not been fulfilled. In [4], in order to determine the location of the receivers, it is assumed that the first source and the first receiver are co-located. This was not considered during the experiments and thus affected the end result.

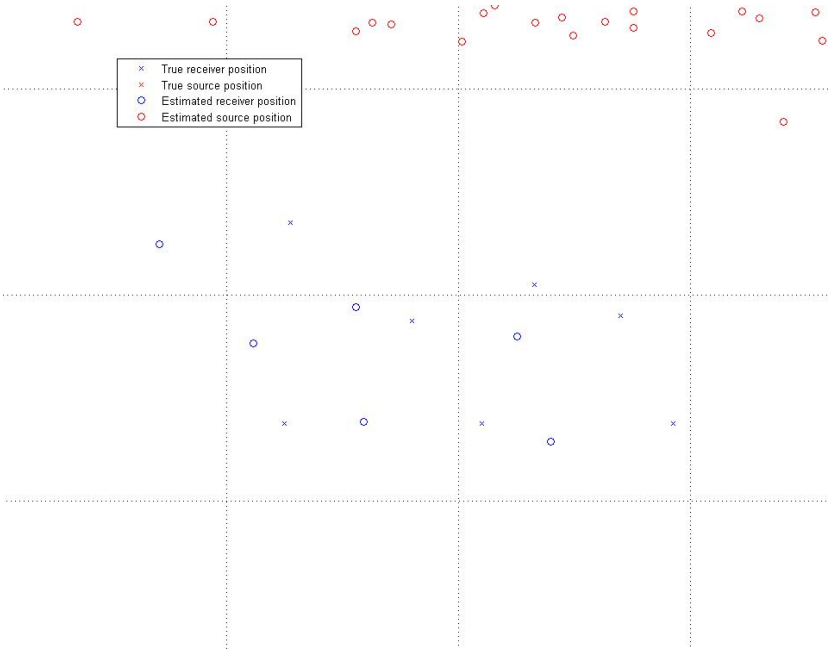



**Figure 6-23:** Receiver estimation closeup. Each grid line is separated by 10 cm

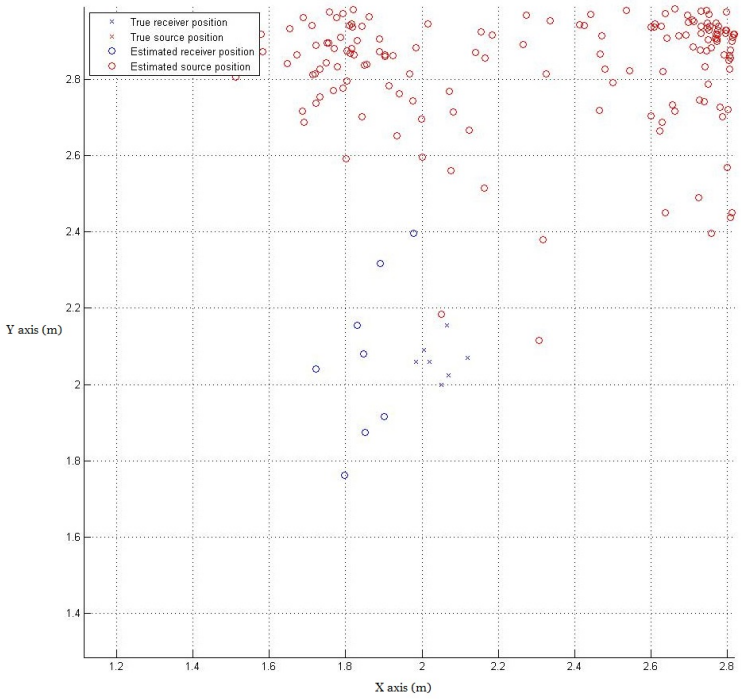



**Figure 6-24:** Receiver estimation example

# Chapter 7

# Conclusions and Future Work

It was stated at the beginning of this work, that determining the onset times and internal delays is crucial for the problem of localization of sources and receivers. Throughout the project, two approaches were studied; one related extending the current framework to time difference of arrival, and the other related to the use of speech as a calibration signal.

There are several conclusions that can be drawn from this project, and there is also room for improvement in this direction:

-Time difference of arrivals between two receivers eliminates the need to determine the onset times. This framework however, proved not be adaptable to current internal delay determination algorithms. As was discussed before, this is due to the cross terms generated in the process, which affect the basic assumptions used in [2] and in [3].

-Although the use of a clicker in [7] provides very accurate results, this can also be inconvenient as was discussed in section 3-0-3. Therefore a readily available calibration signal such as speech is a step forward towards simplifying the localization process. A single person speaking in order to calibrate the system and estimate the onset times, is

-Despite the fact that the error in of the method proposed in [7] is in the mm range and the method proposed here has a degraded performance, only performing in a cm error range, the interest in using speech for the calibration remains an interesting path for further study. The ability to eliminate the need for additional calibration devices, and rely on implicit calibration sources such as speech, is definitely an advantage.

-For future work, it is of interest to consider the variability of the internal delays. The internal delays are related to the state of the receiver (e.g. an Android smartphone), because the same receiver may have different internal delays depending on the state of the device at the time of reception.

-It is of interest to include into further work the effect of reverberations of the calibration signal. That is, take into consideration the effects of the room transfer function, which may also be known a-priori. This will also integrate additional information which can be used to reduce the errors.

# Appendix A

# Glottal Closure Instance Estimation Forward Backward Algorithm (GEFBA)

## A-1 Introduction

GEFBA, which stands for Glottal Closure/Opening Instant Estimation Forward-Backward Algorithm, is a method for speech analysis. It is intended to find the glottal closure instances (GCI) of speech signals. Such speech signals can be voiced, unvoiced or a mix of both (A-1), and GEFBA will output the vector of those GCI from the glottal flow derivative (GFD).

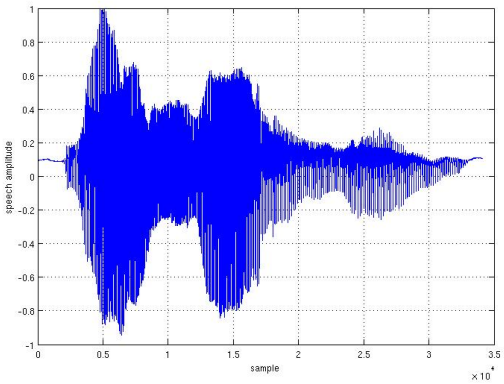



**Figure A-1:** Speech signal sample

The glottal flow of a speech signal represents the velocity of the airflow as it passes through the glottis. This airflow is quasi-periodical due to the glottis opening and closing ([12]) as can be seen in figure (A-2)

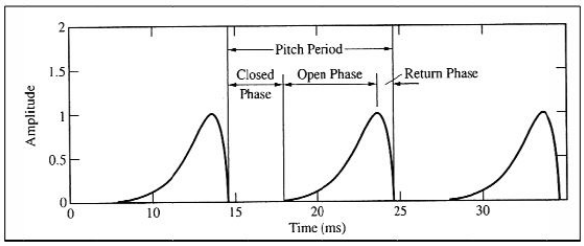



**Figure A-2:** Glottal flow

The GF then passes through a series of filters which model the vocal tract. The output is then voiced speech. Unvoiced speech is not produced from a glottal flow, but it is rather modelled using white gaussian noise (see figure A-3).

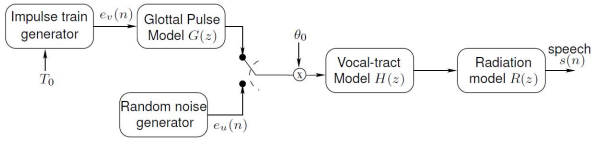



**Figure A-3:** Speech model

## A-2 GEFBA

As stated before, the purpose of GEFBA is to output a vector which contains all the instances at which glottal openings and glottal closures occur. This is done by integrating a voice activity detection step, followed by the identification of the glottal closure and glottal opening instances (GCI and GOI, respectively) by means of the glottal flow derivative (GFD) of the voiced speech signal. The algorithm works in two phases; one which estimates the GFD from the speech signal, and the second one which segments the GFD in voiced and unvoiced parts, and then locates the GCIs and GOIs at those segments.

As mentioned previously, the first step is to calculate the GFD from the speech signal. In order to do this, the speech signal has to be deconvolved with a filter which models the vocal tract and lip radiation ([8]). So, first the speech signal is pre-emphasized with a filter $D(z)$. After performing a linear prediction analysis with 50% overlap to determine the vocal tract and lip radiation, each frame is deconvolved with its corresponding vocal tract filter. The output is the emphasized glottal flow, which is then de-emphasized with $\frac{1}{D(z)}$ and then the GFD is obtained from all of the segments using the overlap-add method.

The next step, which uses the GFD of the speech signal, begins by detecting the maxima and minima, and evaluating those under a set of conditions. So, the first step is to detect a minimum which is named $E_e(i+1)$ in figure A-4 and represents the GCI. The algorithm then looks to the left of this minimum, until it finds the first zero crossing ($t_p(i+1)$ in figure A-4). After this zero crossing is found, it is necessary to find the maximum $E_m(i+1)$ in the $d_e(i+1)$ interval ($d_e(i+1) = t_e(i+1) - t_e(i)$). Once the maximum is found, the algorithm moves to the left in order to find the first zero crossing (see $t_{o1}(i+1)$ in figure A-4). Once this zero crossing if found, the algorithm moves to the right of $t_o(i+1)$ and looks for a $t_{o2})(i+1)$

with an amplitude which is less than $kE_m(i+1)$ $(0 \leq k < 1)$. This point $(t_{o2}(i+1))$ is then designated the GOI.

These two instances (GCI and GOI) are calculated on the whole voiced and unvoiced segments of the speech signal, but the algorithm may actually cause problems at the edges of the voiced speech segments. Then, it is necessary to verify each GCI and GOI with six conditions in order to consider it valid:

$$\begin{aligned}
\mathbf{C1} &= \alpha_1 \mathbf{d_e(i)} < \mathbf{d_e(i \pm 1)} < \alpha_2 \mathbf{d_e(i)} \\
\mathbf{C2} &= \beta_1 \mathbf{d_e(i)} < \mathbf{d_p(i \pm 1)} < \beta_2 \mathbf{d_e(i)} \\
\mathbf{C3} &= \gamma_1 \mathbf{d_c(i)} < \mathbf{d_c(i \pm 1)} < \gamma_2 \mathbf{d_c(i)} \\
\mathbf{C4} &= \delta_1 \mathbf{d_e(i)} < \mathbf{d_o(i \pm 1)} < \delta_2 \mathbf{d_e(i)} \\
\mathbf{C5} &= \eta_1 \mathbf{E_e(i)} < \mathbf{E_e(i \pm 1)} < \eta_2 \mathbf{E_e(i)} \\
\mathbf{C6} &= \zeta_1 \mathbf{E_m(i)} < \mathbf{E_m(i \pm 1)} < \zeta_2 \mathbf{E_m(i)}
\end{aligned} \tag{A-1}$$

where $d_p(i+1) = t_p(i+1) - t_p(i)$, $d_o(i+1) = t_{o2}(i+1) - t_{o2}(i)$ and $d_c(i+1) = t_e(i+1) - t_p(i+1)$.

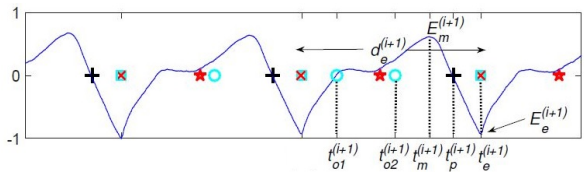



**Figure A-4:** Speech signal sample

The set of control parameters seen in (A-1) are divided into two groups, where $0 < \alpha_1, \beta_1, \gamma_1, \delta_1, \eta_1, \zeta_1 < 1$ and $1 < \alpha_2, \beta_2, \gamma_2, \delta_2, \eta_2, \zeta_2$. When any of the conditions in (A-1) is not met, the GCI and GOI pair are discarded, since it means that part of the GFD corresponds to an unvoiced speech segment or at least a part of the speech segment right at the limit between a voiced and unvoiced block.

The forward-backward nature of the algorithm works one pitch period at a time. This means, that the algorithm moves forward and backwards one pitch period in order to estimate the next set of glottal parameters. Once the GCI candidates are found they are analysed using the conditions in (A-1). The forward movement part of the algorithm is described next, where the search interval is defined as $[t_e(i) + \alpha_1 d_e(i), t_e(i) + \alpha_2 d_e(i)]$:

**MF1**: Search the zero crossings $(t_p(i), t_o(i))$ **MF2**: Find the minimum between two consecutive zero crossings $(min[t_p(i), t_{o1}(i+1)])$. Those minima are the $N$ GCI candidates. **MF3**: Using the GCI found before, calculate the rest of the parameters $(E_m, t_{o2}, d_e(i), t_m(i))$ **MF4**: Discard all GCI and GOI pairs that dont comply with the conditions shown in (A-1). If the resulting number of glottal parameters $(M)$ is reduced to zero, this means that the speech signal analysed is unvoiced. **MF5**: Use the remaining $M$ sets as the next set of glottal parameters.

The backwars movement part is the same as the forward movement, but the interval is now $[t_e(i) - \alpha_2 d_e(i), t_e(i) - \alpha_1 d_e(i)]$.

The second phase of the algorithm is a more relaxed search, which first identifies the voiced segments and re-estimates the glottal parameters on a frame approach. In step 1 of this phase (see A-5), the algorithm processes a frame and classifies it in one of two: highly voiced, or not, based on whether or not it fulfils the conditions in (A-1). In the end, the first step outputs the beginning and end of voiced segments.

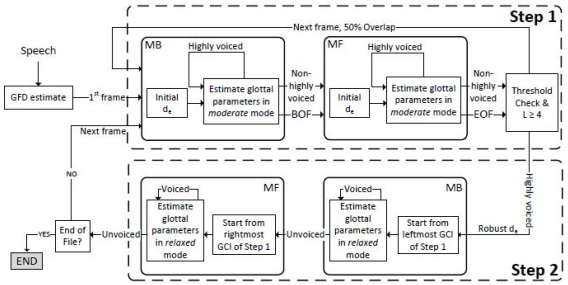



**Figure A-5:** Block diagram of the method (from [1])

With this information, step 2 fills those "gaps" in a more relaxed mode for the condition parameters (as in (A-1)). Finally, the algorithm produces a list of the GCIs and GOIs of the entire speech signal. From these glottal parameters, it is possible to calculate the pitch periods.

# Bibliography


[1] R. H. Andreas Koutrovelis and R. Hendriks, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015.

[2] R. Heusdens and N. Gaubitch, "Time-delay estimation for toa-based localization of multiple sensors," *ICASSP*, 2014.

[3] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," *ICASSP*, 2008.

[4] M. B. Marco Crocco, Alessio del Blue and V. Murino, "A closed form solution to the microphone position self-calibration problem," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2012.

[5] S. Thrun, "Affine structure from sound," *NIPS proceedings*, 2005.

[6] K. A. Yubin Kuang, "Stratified sensor network self calibration from tdoa measurements," *European Signal Processing Conference (EUSIPCO)*, 2014.

[7] B. K. Richard Heusdens, Nikolay Gaubitch, "Calibratio of distributed sound acquisition systems using toa measurements from a moving acoustic source," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.

[8] A. G. J.D.Markel, *Linear Prediction of Speech*. Berlin: Springer-Verlag, 1976.

[9] T. L. Feng Huang, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," *IEEE transactions on Audio, Speech and Language Processing*, 2012.

[10] S. Schlect and E. Habets, "Time-varying feedback matrices in feedback delay networks and their application in artificial reverberation," *Acoustical Society of America*, 2014.

[11] T. L. et al., "Splitting the unit delay," *IEEE Signal Processing Magazine*, 1996.

[12] G. Kafentzis, "On the glottal flow derivative waveform and its properties," tech. rep., University of Crete, 2008.