Chapter 2

Prerequisites

For the whistle sound source localization with multiple robots, some sequential steps needs to interact for a final result. To work with the implementations in chapter 3, the fundamentals are introduced in this chapter on a general basis.

As this work specializes on the localization of a whistle sound, this sound pattern must be detected at first. This was already done by previous work [1]. Hence, the flow of the whistle detection is only briefly explained in section 2.1. The position of the sound source is determined by combining separate direction results on the single robots. Utilizing the TDOA information referenced in section 2.2 between pairs of the Naos' microphones, every robot produces a sound direction ray which are fed into the team decision filter. As mentioned in chapter 1, different methods exist to identify the TDOA and are terms of content in sections 2.3 to 2.5. Due to the low resolution arising from the sample rate and small distance between the microphones, a subsample estimation is stated in section 2.6. One of the most significant factors is the selection of the signal frame that is going to be used for the direction determination. In order to provide a correct and stable decision process, different approaches are considered. In all cases, the focus is on finding the signal starting index and are listed in section 2.7. After all, the results of the individual robots are filtered by assuming gaussian distribution to produce a sound source position. Section 2.8 describes the formulas for Bayesian Updating.

2.1 Whistle Signal

In this work the localization of a whistle sound source is to be to the fore. Detection of the whistle is done in frequency domain by assuming the whistle sound to be higher than 2kHz and lower than 4kHz. By comparing the mean of the signal between this band with the overall mean of the received signal, a peak arising around the whistle frequency can be detected. For the whistle detection, only one channel of the robot is used and the mean of the whistle band must exceed the threshold multiple cycles in a row. If the team takes action due to the detected signal on individual robots is a team decision.

Further on for this work, the mathematical model of a received whistle signal at one microphone sensor is defined as

$$x_i(t) = s_i(t) + n_i(t) \text{ for } i \in \{0, 1, 2, 3\} \tag{2.1}$$

where $s(t)$ represents the signal and $n(t)$ noise. Both are assumed as real, jointly stationary random processes.

2.2 Time Difference Of Arrival

The direction of a signal source γ' can be determined by the time delay of the received signal. Calculations for the direction of the sound source can be done with a geometrical approach like in [2]. Figure 2.1 illustrates how a the delay D is introduced by the direction angle of the sound source relative to a vector between channels 0 and 1. Here, the signal arrives first at channel 0. Ideally, the same signal values are measured at channel 1 with delay D. If the delay is zero, the signal is perpendicular to the channels vector. It's value can be D_{max} maximally which in that case delivers the result of the source direction vector being aligned to the channels vector. It is assumed that the distance from the sensors to the sound source is significantly large so that the signal waves proceed parallel which is a necessary criterion for the approach to be valid.

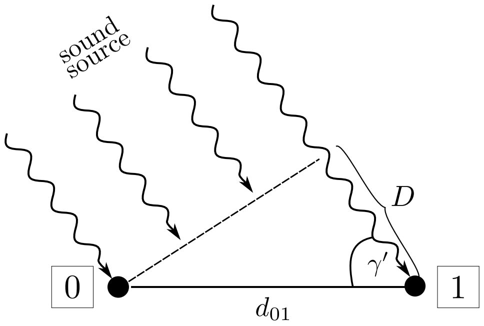

Figure 2.1: *Illustration of TDOA principle.*

Specifying the speed of sound c_s being 343m/s in air, the angle γ' can be defined as

$$\gamma' = cos^{-1}(\frac{|D|}{D_{max}}) \tag{2.2a}$$

with

$$D_{max} = \frac{f_s \cdot d_{01}}{c_s} \tag{2.2b}$$

where f_s is the sampling rate and d_{01} is the distance between both channels. Not to forget is the ambiguity of the result by observing two channels only. Having a look at fig. 2.1 once more and assuming that the sound source is positioned in front in this case, the same delay can be the result of a sound source from behind. For quick understanding, one can find an illustration of the second possible source directions in fig. A.1.

With the definition of a whistle signal as stated in eq. (2.1), the microphone sensors *channel*$_0$ and *channel*$_1$ will output

$$x_0(t) = s(t) + n_0(t) \tag{2.3a}$$

$$x_1(t) = \alpha s(t - D) + n_1(t). \tag{2.3b}$$

Again, D is the delay of x_1 relative to x_0 for what is looked for. As introduced in chapter 1, different methods to detect this delay were implemented and evaluated in this work. In the following sections, the theoretical background of these will be explained in detail.

2.3 Cross Correlation

The Cross Correlation (CC) provides information about the similarity of two signals. Thus, the delay of one signal can be detected where the CC $R_{12}(t)$ is largest. In time domain, the CC of two signals x_0 and x_1 is denoted as

$$R_{x_0x_1}(t) = x_0(t) \circledast x_1(t) = \int_{-\infty}^{+\infty} x_0(\tau - t)x_1(\tau)d\tau. \tag{2.4}$$

Considering the frequency domain, the function can be transformed into

$$\mathcal{F}[R_{12}(t)] = G_{x_0x_1}(f) = X_0^*(f)X_1(f) \tag{2.5}$$

with $\mathcal{F}[x_i(t)] = X_i(f)$ and $X_i^*(f)$ indicating the conjugate complex form. However, the finite observation time of the received signal corrupts the fourier transform [3] and noise of sensors may introduce false peaks in the CC [4]. In frequency domain, the signals $x_0(t)$ and $x_1(t)$ from eq. (2.3) can be expressed as

$$X_0(f) = S(f) + N_0(f) \tag{2.6a}$$

$$X_1(f) = \alpha S(f)e^{-j2\pi fD} + N_1(f). \tag{2.6b}$$

Thus, the CC is

$$G_{x_0x_1}(f) = \alpha|S(f)|^2e^{-j2\pi fD} + N_0^*(f)N_1(f) + S^*(f)N_1(f) + \alpha S(f)e^{-j2\pi fD}N_0^*(f) \tag{2.7a}$$

which will be shortened as

$$G_{x_0x_1}(f) = \alpha\phi_s(f)e^{-j2\pi fD} + \phi_n(f) + \phi_c(f) \tag{2.7b}$$

where

$$\phi_s(f) = |S(f)|^2 \tag{2.7c}$$

$$\phi_n(f) = N_0^*(f)N_1(f) \tag{2.7d}$$

$$\phi_c(f) = S^*(f)N_1(f) + \alpha S(f)e^{-j2\pi fD}N_0^*(f). \tag{2.7e}$$

Considering the ideal case where $s(t)$, $n_0(t)$ and $n_1(t)$ are uncorrelated, the terms ϕ_c and ϕ_n disappear and the CC results in

$$R_{12}(t) = \mathcal{F}^{-1}[\alpha\phi_s(f)e^{-j2\pi fD}] = \alpha\mathcal{F}^{-1}[\phi_s(f)] \circledast \delta(t - D). \tag{2.8}$$

The CC gives insight about the similarity of two signals and at peak, they are most alike. Thus, the shift between the zero index and the peak is the resulting delay. In general, ϕ_c and ϕ_n can neither be neglected nor assumed as uncorrelated to the signal [5], so that they introduce inaccuracies and errors.

Figure 2.2 is the outcome of two similar, but shifted sine signals with 3kHz and normally distributed noise. As the second signal is delayed by 10 samples, the peak can be detected where $shift = 10$. The example signals are attached in fig. A.3. One disadvantage of this technique is that for periodic signals the CC also is periodic and the peak is not always easily detectable. Noise and inaccuracies of the Fast Fourier Transform (FFT) then may influence the result what can make the peak unobvious [6].

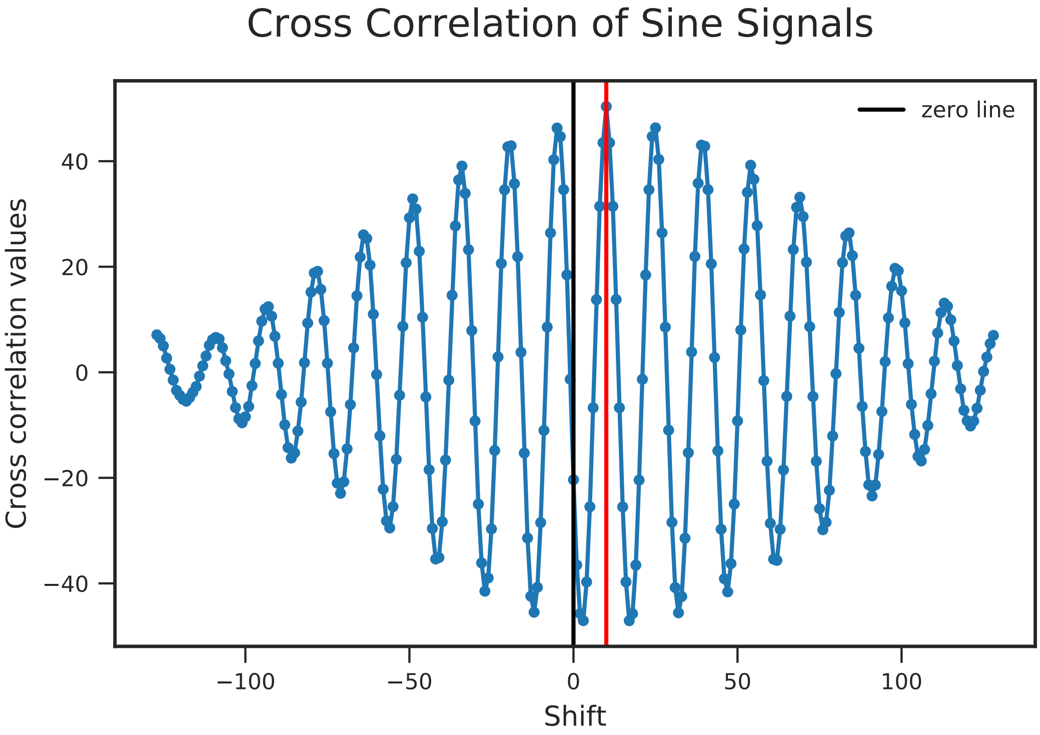

Figure 2.2: *Cross correlation of two generated sine signals with 3*kHz*.*

2.4 Generalized Cross Correlation

As depict in section 2.3, the CC can bring some error sources in the context of incorrect delay results and inaccuracy. Improvements were done in research by introducing prefilters for the signals which is equal to general frequency weighting as stated in [3]. With certain weightings $H_i(f)$ prior to the CC, the peak detection can be rectified by improving the relation between peak and noise or enhancing the accuracy [4]. Figure 2.3 illustrates the process of a Generalized Cross Correlation (GCC) with both filters combined as $\Psi(f) = H_0^*(f)H_1$. The figure in appendix A.2 represents the GCC with $H_i(f)$. After transforming the signals $x_i(f)$ into frequency domain, the cross correlated signals are multiplied with the weighting $\Psi(f)$ and transformed back into time domain. The subsequent steps are similar to the CC.

Thus, the GCC is declared as

$$R^{(g)}_{x_0x_1}(t) = \int_{-\infty}^{+\infty} \Psi(f)G_{x_0x_1}(f)e^{j2\pi ft}df. \tag{2.9a}$$

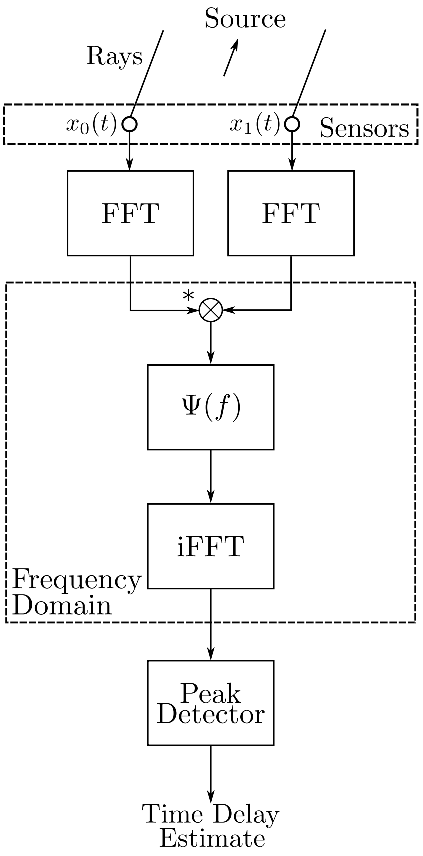

Figure 2.3: *Generalized cross correlation for time delay estimation*

Written-out it is visible how the choice of $\Psi(f)$ impacts the individual segments of eq. (2.7) as

$$\begin{aligned} R^{(g)}_{x_0x_1}(f) = \mathcal{F}^{-1}[\Psi(f)\alpha\phi_s(f)e^{-j2\pi fD}] \\ + \mathcal{F}^{-1}[\Psi(f)\phi_n(f)] + \mathcal{F}^{-1}[\Psi(f)\phi_c(f)]. \end{aligned} \tag{2.9b}$$

Several variants of the weighting were designed by various researchers with different criteria. They have in common, that they take the characteristics of the received signals into account. Some favor one of both signals, some are designed to suppress the noise and other focus to sharpen the peak as contrasted in [3]. The characteristics of the GCC with Phase Transform (PHAT) most appealed to the task in this work and is chosen as weighting function.

2.4.1 The Phase Transform (PHAT)

The PHAT weighting is known as

$$\Psi^{(P)}(f) = \frac{1}{|G_{x_0x_1}(f)|}. \tag{2.10}$$

For the ideal case that ϕ_n and ϕ_c are nonexistent due to non-correlation, the the GCC results in

$$R^{(p)}_{x_0x_1}(t) = \mathcal{F}^{-1}[\frac{\alpha|S(f)|^2e^{-j2\pi fD}}{|G_{x_0x_1}(f)|}] = \delta(t-D) \tag{2.11}$$

because $|G_{x_0x_1}(f)| = \alpha|S(f)|^2$. This filter is used regularly in research, due to the characteristic of sharpening the peak what leads in high accuracy.

Figure 2.4 shows the result of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) algorithm with a generated simple Hann-windowed signal as input. Both signals are 3kHz sine signals, whereby the second signal is shifted by 10 samples. The signals are similar to the ones used in section 2.3 and are plotted in fig. A.3. Compared to the CC in section 2.3, the sharp peak is distinct. However, [3] states about the lower robustness of this algorithm.

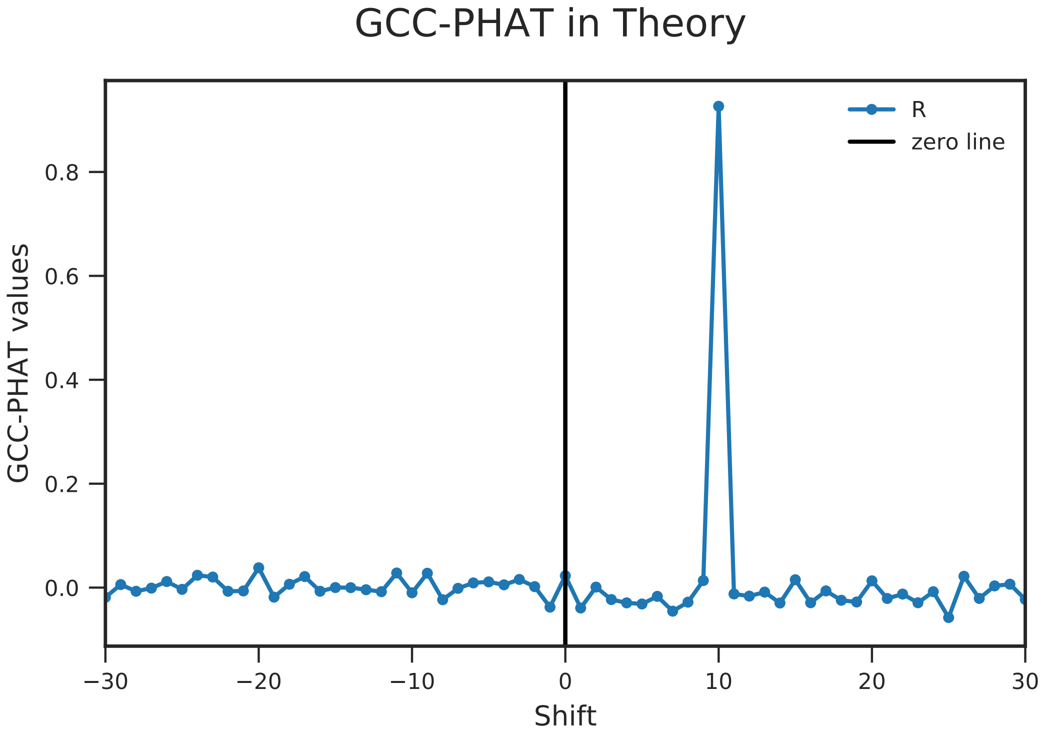

Figure 2.4: *Generalized Cross Correlation with PHAT weighting of two generated 3*kHz *sine signals.*

2.5 Signal Phase Difference

With a different approach to the correlation methods, the TDOA can be detected by observing the phase of one certain frequency. The phase of a signal frequency is easily computable in frequency domain with

$$\phi(f) = tan^{-1}(\frac{imag(X(f))}{real(X(f))}). \tag{2.12}$$

With the difference of the phases of two channel, the delay is defined as

$$D = \frac{\Delta\phi \cdot c_s}{2\pi \cdot f}. \tag{2.13}$$

From that, the direction angle calculation of eq. (2.2a) can be followed.

2.6 Subsample Shift

Considering the case that the sample frequency f_s is set to 44.1kHz and the sound speed is 343m/s, the maximal number of samples between the rear channels is 14. Other neighboring pairs have even less maximal sample differences. This leads to a very low resolution of the direction angle which can be circumvented by either setting a higher resolution or interpolation.

Quadratic interpolation is a well known technique to obtain a floating number shift from a correlation. For this, a parabola $y(x) = a(x-p)^2 + b$ is fitted into the three values of R around the peak of the CC and the peak of the parabola is taken as the more accurate delay. Thus, the subsample delay D_{sub} depends on the maximal value of the correlation y_m and it's previous one y_{m-1} and the next value y_{m+1}. Substituting known values and derivations into the parabola function, the subsample delay is defined as

$$D_{sub} = \frac{y_{m-1} - y_{m+1}}{2 \cdot (y_{m-1} - 2y_m + y_{m+1})} \tag{2.14}$$

like in [7]. Figure 2.5 illustrates the CC of two generated sine signals with 3kHz. The second signal is shifted by $\frac{\pi}{3}$ which are 2.449 samples for a sample rate of 44.1kHz. As the plot shows, the peak of the parabola can be determined at an index of 2.446 by quadratic interpolation. In research there are efforts in finding a better approximation function than the quadratic as stated in [8] but these are not discussed in greater detail here.

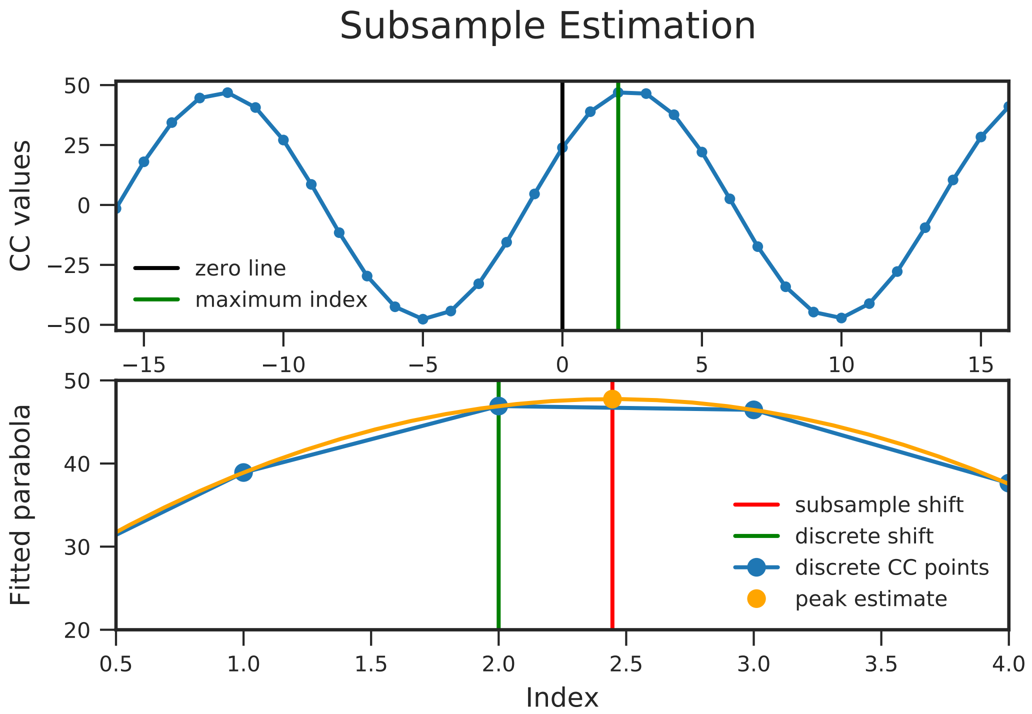

Figure 2.5: *Explanation example of the subsample shift estimation using parabolic interpolation.*

2.7 Signal Start Detection

One focus of the whistle signal localization is the correct choice of the signal frame, with which the TDOA calculation is done. Assuming that the clearest signal without reverberation and with minimal multipath propagated subsignals is at the start of a sound signal, the frame to investigate is chosen to be at the beginning of a whistle sound.

By knowing the frequency band of a whistle signal, the start can be detected where these frequencies dominate. Using this indicator only does not always give the desired accuracy, that is why different methods are investigated in this work. In the next subsections, signal start detection using short time energy, zero crossing rate and spectral entropy are subject of discussion. Also, the methods require various computational power. According to the circumstances, the most suitable approach can be chosen. Another point is, that robustness can be increased by considering these methods in combination. As a latter, the consensus of the single methods can be passed as information about the certainty of the computed direction result.

2.7.1 Short Time Energy and Zero Crossing Rate

A common method in signal start and endpoint detection is the evaluation of the Short Time Energy (STE) and Zero Crossing Rate (ZCR).

2.7.1.1 Short Time Energy

The energy

$$E = \sum_{n=1}^{N} E_s(n) \tag{2.15}$$

with the energy spectral density

$$E_s(n) = |X(n)|^2 \tag{2.16}$$

of signal frames with length N are expected to be higher than noise frames and therefore, noise and signal can be distinguished according to [9]. A threshold needs to be specified appropriately dependent on the environment.

2.7.1.2 Zero Crossing Rate

The ZCR of one frame Z needs small computational effort in order to identify a periodic signal in time domain. Its formula is

$$Z = \sum_{k=2}^{N} |sgn(x(k)) - sgn(x(k-1))| \tag{2.17}$$

with the sign function

$$sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

for a discrete signal $x(k)$ of a frame with length N [9]. A higher ZCR is an indication for a periodic signal. To detect the signal start, a threshold is determined dynamically. The ZCR mean of frames which are known to be noise only are averaged with the mean of those frames, that include the whistle signal. The signal start is detected at the point in time, where the ZCR exceeds this threshold.

2.7.2 Spectral Entropy

Entropy provides information about the disorder of a system. From this, one can derive that noise has a high entropy compared to a whistle sound, which is a highly structured sound signal and a high amount of information accordingly. The spectral entropy of a signal is determined by normalizing the Probability Density Function (PDF) over all frequency components as described in [10]. When $X(n)$ is the Discrete Fourier Transform (DFT) of the sampled signal $x(k)$, the PDF is

$$P(n) = \frac{E_s(n)}{E} \tag{2.18}$$

with eq. (2.16) as the spectral energy density function for $E_s(n)$ and E as the energy. Finally, the spectral entropy results in

$$H = -\sum_{n=1}^{N} P(n) log_2 P(n). \tag{2.19}$$

Utilizing some a priori knowledge about the signal, the entropy estimation can be improved. In this work, the frequency of a whistle sound is known to be between 2kHz and 4kHz from [1], Thus, only the frequency components in the whistle range is considered. Differentiating between noise samples where no signal is present and signal frames, a dynamic threshold can be set to detect the signal start point.

2.8 Bayesian Updating

Assuming gaussian distribution of the single robot results, the multi-agent decision process is done by Bayesian Updating. One dimensional probability density functions of states with variance σ^2 and mean μ are described as

$$\mathcal{N}(x,\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\sigma)^2}{2\sigma^2}}. \tag{2.20}$$

Having two values μ_0 and μ_1 with their variances, the result μ' of the combination of both is

$$\mathcal{N}(x,\sigma',\mu') = \mathcal{N}(x,\sigma'_0,\mu'_0) \cdot \mathcal{N}(x,\sigma'_1,\mu'_1). \tag{2.21}$$

By substitution and conversion, μ' and σ'^2 can be formulated to

$$\mu' = \mu_0 + \frac{\sigma_0^2(\mu_1 - \mu_0)}{\sigma_0^2 + \sigma_1^2} \tag{2.22a}$$

$$\sigma'^2 = \sigma_0^2 - \frac{\sigma_0^4}{\sigma_0^2 + \sigma_1^2}. \tag{2.22b}$$