Broadband sound localization: Two channel SRP-PHAT based direction of arrival estimation on SHARK DSP

(Algorithms in Signal Processors)

Denis Nadein, ic05dn6@student.lth.se

March 25, 2015

Contents

Abstract

A dual channel direction of arrival estimation algorithm based on the SRP-PHAT model is implemented on the SHARC ADSP-21262 platform with initial algorithm evaluation, behavioral modeling and finally an embedded implementation targeting the platform. The key mechanism is based on generalized cross correlation and, by extension, steered response power with phase transform algorithms. The two-channel array based implementation is able to indicate approximate direction of a sound source within a half circle of the microphone array. The quadrant where the source is localized is indicated by means of a distinct sinusoidal signal sent through headphone output. Furthermore, the amplitude of the sinusoidal is modulated by the derived phase angle deviation in the quadrant. Calibration is performed to establish an approximate origin point when the sound source is straight in front of the array.

1 Introduction

The purpose of this document is to describe the implementation of a two channel direction of arrival estimation unit and the underlying theory. The theory and the workings of the signal processing framework is tested in Matlab prior to implementation on a SHARK based DSP platform in the embedded subset of the C programming language. The goal is a working prototype that is able to derive approximate direction of a dominant (loudest) sound source in local environment, with the specified active area defined as the front half circle of the microphone array. The unit should be able to determine the origin quadrant of the sound source and approximate deviation from the normal in the arrays uptake in respective quadrant with minimal error.

2 Implementation

Derived from work presented in the doctoral dissertation of Mikael Swartling on direction of arrival estimation and localization of multiple speech sources in enclosed environments [1]. The signal processing theory used for the implementation of the direction of arrival determining algorithm is based on *Generalized Cross-Correlation* and *Steered Response Power with Phase Transform.*

2.1 Microphone array geometry

The calculation is done by reading the two channels in frequency domain and converting the phase difference of the incoming signal to time difference in frequency domain. Since the speed of sound can be approximated, so can the angle of arrival relative the two microphones (channels), as shown in figure 1. The sound signal reaches the microphones with a slight delay and this information is used to calculate the approximate direction of signal source. The equation below is a trigonometric relationship between time difference of sounds arrival τ, distance between the microphones d and the angle θ between a microphones normal and the incoming sound wave-front.

$$\tau = dsin\theta \tag{1}$$

The time to frequency conversion is done by means of *Fast Fourier Transform* and to avoid *spatial* aliasing, the phase difference between signals is constrained to be less than π. I.e., the shift in time of sounds arrival between each microphone can not be more than the corresponding time shift. With F_S being the sampling frequency, the speed of sound c and the distance between the microphones d, maximum time delay τ_{max} can be calculated:

$$d = \frac{\tau_{max}}{F_s/c} \iff \tau_{max} = d \cdot \frac{F_s}{c} \tag{2}$$

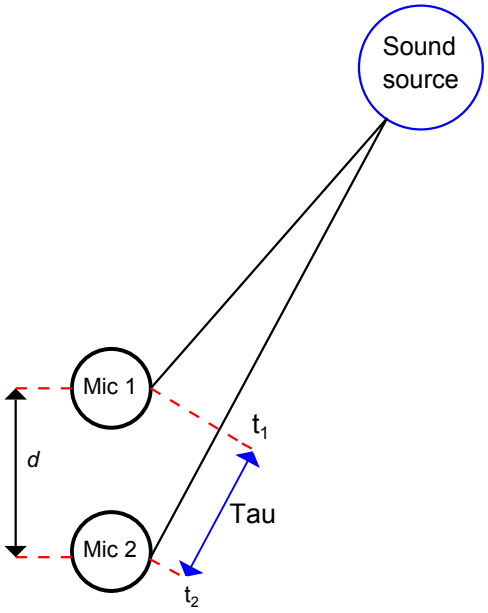

Figure 1: Illustration of direction of arrival estimation with a two microphone linear array. The distance between the microphones is d. Time of the sounds wave front arrival at the first and second microphone are t_1 and t_2, respectively. Tau (τ) is the time difference of arrival, i.e. difference of t_1 and t_2.

2.2 Generalized Cross Correlation

The two channels are represented in time domain as $x_1(t_2)$ and $x_1(t_2)$, with t_1 and t_2 being the time of signals arrival at the respective microphone. This can also be expressed in terms of relative time of arrival: $s(t)$ and $s(t-\tau)$ with τ being the difference in the sounds time of arrival between the two microphones. The signal of interest is therefore the same as reference but shifted in time with τ, meaning it can be translated (as demonstrated by equation 3 below) to a phase shift in frequency domain. The additional variable ω represents the frequency of the signal.

$$\begin{aligned} X_1(\omega) &= FFT(s(t)) \\ X_2(\omega) &= FFT(s(t-\tau)) = X_1(\omega) \cdot e^{-j\omega\tau} \end{aligned} \tag{3}$$

The *cross power spectrum* function $G_{1,2}(\omega)$ of the two signals shown in equation 3 leads to information on the phase offset by comparing where power peaks co-incide. After stating that the complex conjugate of $X_2(\omega)$ can be rewritten in terms of the original signal as $X_2^*(\omega) = X_1^*(\omega) \cdot e^{j\omega\tau}$, the cross power spectrum is defined as following:

$$G_{1,2}(\omega) = X_1(\omega) \cdot X_2^*(\omega) \tag{4}$$

Frequency ω is in radians, $X_2^*(\omega)$ is the conjugate of $X_2(\omega)$ and τ is the phase difference between X_1 and X_2. Note that so far, the calculation is only producing one quantity as output, that of the power of the spectrum with regard to frequency ω. Since the goal is to acquire an estimation of τ when it is not known, the equation 4 needs to be extended.

Vector $\overline{\tau}$ is introduced, with individual elements referred to as $\hat{\tau}$. It is a set of values based on τ_{max} in equation 2 which are dividing the microphone array uptake in a set of discrete angle values. These values are used to *test* for the correct τ by maximizing following function, called the *Generalized Cross Correlation*:

$$R(\hat{\tau}) = \frac{1}{2\pi} \sum_{k=1}^{N/2} \Psi_{1,2}(k) \cdot G_{1,2}(k) \cdot e^{\frac{-j2\pi k}{N}\hat{\tau}} \tag{5}$$

Frequency earlier referred to as ω is converted to discrete domain tap number k, to be more suitable for in-hardware calculations, N is the block size used to perform the Discrete Fourier Transform on the DSP and $\hat{\tau}$ is an element of the test vector. For each individual value $\hat{\tau}$ in vector $\overline{\tau}$, equation 5 is calculated and the value $\hat{\tau}$ that generates the highest $R(\hat{\tau})$ is the one that with high probability is closest to the actual τ. The factor $\Psi_{1,2}(k)$ is what is referred to as the *phase transform processor* and contributes to mitigate effects of reverberation in the environment [1] and is defined in context of equation 5 like following:

$$\Psi_{1,2}(k) = \frac{1}{|X_1(k) X_2^*(k)|} \tag{6}$$

2.3 Steered Response Power with Phase Transform

The most probable delay between two microphones is calculated by checking where the cross correlation between the two signals is the highest (meaning that the probability that the comparison is being done between the same signal at a different time is highest). The equation 7 below summarizes the main principle of the SRP-PHAT algorithm:

$$\tau_{\delta_{1,2}} = \arg\max_{\overline{\tau}} (R(\hat{\tau})) \tag{7}$$

Where $R(\hat{\tau})$ is the GCC function shown in 5. The derived value of τ_{max} in equation 2 is used as a guideline for construction of the vector $\overline{\tau}$. The individual values of the test vector are tested to determine the value that maximizes $R(\hat{\tau})$ for each processed block, i.e. over all Discrete Fourier Transform taps. The $\hat{\tau}$ is the individual constant passed during each iteration, until the length of the test vector has been processed. Sampling frequency and block size used in the DSP implementation are taken into consideration. The resulting output $\tau_{delta_{1,2}}$ is the estimated delay between signals arrival at the two microphones. The accuracy of the result is dependent on choice of values in the test vector.

2.4 Model

The resulting model for the *Steered power response with phase transform* algorithm for direction of arrival estimation 8 is summarized below in equation 8:

$$\tau_{\delta_{1,2}} = \arg\max_{\overline{\tau}} \frac{1}{2\pi} \sum_{k=1}^{N/2} \Psi_{1,2}(k) \cdot G_{1,2}(k) \cdot e^{\frac{-j2\pi k}{N}\hat{\tau}} \tag{8}$$

The block size used in the DSP is $N = 512$. For each sampled block, the algorithm evaluates the result for each element in the test vector $\overline{\tau}$. Only half of the total number of taps N need to be sweeped, due to the symmetrical characteristic of the Fourier Transform. The summation range can additionally have the k offset `+10` and `-10` at the beginning and end, respectively, to decrease the influence of spectral leakage. The test vector $\overline{\tau}$ is symmetrical at zero, to be able to localize source origin from both quadrants in its uptake radius. Fundamental distance limits of the microphones and the sounds propagation speed are taken into consideration as described earlier. The $\hat{\tau}$ that generates the maximum value of 8 is the probable representation of the signals phase difference. Based on this fact, the direction of arrival of one signal relative to the other is calculated.

For testing the model, a sine wave is used as the reference signal (first channel) and a phase shifted version of the same signal to represent the second channel. Static noise is added to account for background noise in the real environment.

2.5 C

The hardware implementation is written in C, using the framework provided by the libraries available in the Visual DSP++ environment and supplied as part of course material.

A 512 sample large block size is used in the DSP, upon which a real DFT with the same number of taps is performed, with the sampling frequency used being `16` kHz. A hamming window is subsequently applied to reduce the spectral leakage. To provide user feedback, a sine wave is used as output, with the amplitude modulated by the phase difference as acquired by the *SRP-PHAT* algorithm. The frequency of the sine wave is different for each quadrant. The output from 8 is calibrated for the microphone array to center its origin when the sound source is approximately straight in front of the array, at which point the output is only silence.

3 Problems

The selection of a vector $\overline{\tau}$, proved to be a challenge due to requiring a large amount of trial-and-error. Ultimately, a slightly larger value than the calculated τ_{max} proved satisfactory to achieve a functional in-hardware implementation.

The localization is achieved in the complete uptake radius but with better sensitivity in one quadrant. Possible reasons for this is the imperfect array calibration, along with not optimal values in the vector $\overline{\tau}$ used to test for maximum cross-correlation output during selection of best candidate $\tau_{\delta 1,2}$. The usage of only two microphones is naturally also a limiting factor for achieved accuracy within the quadrants. Additionally, appliance of a hamming window function was required, prior to which the accuracy of the angle estimation was significantly lower supposedly due to spectral leakage present in the raw FFT data of the input signals.

4 Discussion & Conclusion

As expected, work with real sound waves in real environment proved to be a challenge when compared to the model level setting. Ideally, a more robust work during the modeling is required. Particularly with the test-input signals. The presence of the spectral leakage phenomenon and subsequent requirement of windowing of processed data is evident during attempts to achieve a functional design in a real environment. Furthermore, the details of how the hardware handles its signals internally and its own noise figures can not always be derived or require considerable further analysis and familiarity, alternatively signal analysis of complete chain from the microphone array to DSP output. Although, the results in regard to broadband sound source localization can be considered satisfactory with only two microphones used in the array. The prototype is able to provide clear and relatively accurate real time feedback on direction of a white noise source during a sweep in its uptake radius.

Bibliography

[1] Mikael Swartling. *Direction of arrival estimation and localization of multiple speech sources in enclosed environments*. School of Engineering, Blekinge Institute of Technology, Karlskrona, 2012. Diss. Karlskrona : Blekinge tekniska högskola, 2012.