AUDIO SOURCE LOCATION FOR A DIGITAL TV-DIRECTOR

Feico W. Dillema, Paul J.M. Havinga, Paul Sijben, Gerard J.M. Smit

University of Twente, department of Computer Science
P.O. Box 217, 7500 AE Enschede, the Netherlands
{dillema, havinga, sijben, smit}@cs.utwente.nl

Abstract

Three algorithms are presented for location of audio sources using standard workstations and a minimal amount of resources. The audio source location is based on time-delay estimation. The algorithms use general human speech properties and straightforward heuristics on human speaker behaviour to acquire accurate and efficient estimation of delays.

1 INTRODUCTION

Audio source location is studied as part of the Pegasus project[1] at the University of Twente. The problem deals with locating and tracking human speakers. The Pegasus project aims at providing general-purpose operating-system support for distributed multimedia applications. Several multimedia applications are under development and their use is to reveal requirements of multimedia for the architecture and implementation of the system. One of the applications is a digital TV-director (Mullender 1994). This application will control cameras and light settings during meetings and conferences. The camera's are mounted on pan-tilt devices that can be controlled from a workstation. In order to aim cameras and spotlights at speakers automatically, the application needs a way to locate audio sources.

Audio source location has been investigated for a number of purposes and applications. Most of these applications use arrays of microphones. Such an array has the potential of producing a beam-formed combination of its received signals, supplying its application with a high-quality signal from a particular audio source. These systems have been developed and used for tele-conferencing, speech recognition, speech acquisition and other applications requiring speech input (Brandstein and Silverman 1993; Brandstein et al. 1995; Omologo and Svaizer 1996). The basic component of an audio source location system is the time-delay estimator which determines the relative time-delay between signals received by two microphones. Traditionally, correlation techniques have been used in designing such a time-delay estimator. Most existing audio location systems are based on maximizing the cross-correlation function of signals from separate receivers using dedicated hardware and/or Digital Signal Processors.

This article describes three approaches of the audio source location problem using a minimal amount of resources. This implies a minimum number of microphones and a low algorithmic complexity. Dedicated hardware is in our context only acceptable when it is simple and low cost. General human speech properties are used in the design of the algorithms in order to make audio source location based on time-delay estimation feasible.

The techniques described here for audio source location show similarities to techniques used in radar and sonar. One of the major differences, however, with the radar/sonar setting is that radar/sonar receivers deal with the detection of a priori known signals (the signals are transmitted by the radar system, and thus are known), while the exact nature of the received signals is unknown in the audio source location setting. Therefore, audio location based on the well-known detection and estimation techniques from radar technology will not perform very well without adjustments.

The three algorithms presented in this paper were developed in the order as presented in the paper. Results and knowledge gained were used in the successive algorithms. However, it is difficult to compare the algorithms, because each algorithm uses a different approach and has its own characteristics and resource requirements. The first two approaches use a standard workstation on which the algorithm is executed in real-time. The third algorithm takes a different approach: its uses inexpensive dedicated hardware. The remainder of this paper describes the algorithms and their design in more detail.

2 BACKGROUND

2.1 Basic algorithm

The sampled signals from two microphones can be used to determine one coordinate of an audio source by estimating the time-delay θ between the two received signals. Let d be equal to the distance of the source to the microphone pair and d_m be equal to the distance between the microphones. The angle of the

1. The Pegasus Project is a project of the Universities of Twente and Cambridge supported by the European Communities Esprit Programme through BRA project 6586.

audio source to the middle of the microphone pairs is denoted by α. From the time-delay θ, the distance-difference δ can be calculated according $\delta = c.\theta$, with c denoting the propagation speed of sound through air[1]. This view is depicted in Figure 1. Each microphone

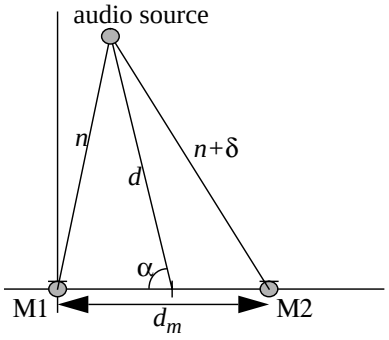

Figure 1: Basic setting employing two microphones

added to this initial setup would then enable another coordinate to be determined. So, a minimal setup for location in the plane contains three receivers, while location in 3D-space requires a minimum of four microphones. By estimating δ in Figure 1 the valid locations for an audio source are reduced to those defined by a hyperbola (see Figure 2). In this paper we

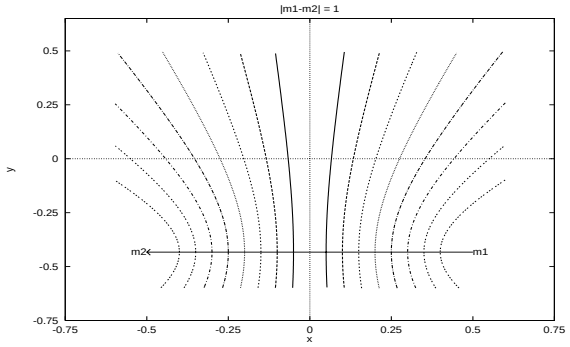

Figure 2: Hyperbolae corresponding to δ's ranging from 0.1 to 0.8 m

focus on experiments using only two microphones, thus restricting the possible locations of an audio source to a hyperbola in a plane. This can be extended to more dimensions with more microphones using similar techniques.

2.2 Environment

The *target environment* will be an enclosed space, typically an office or meeting room. We assume that reflections of audio sources and reverberant noise are negligible, and therefore the sound received by the microphones originates directly from its sources. However, the speaker to be located is not necessarily the only present audio source in the room. Sound irrelevant for location purposes, and therefore noise to the system, can and probably will be present. Noises to expect in a typical office environment are for instance noise generated by electrical devices and the more unpredictable sounds like the ones originating from closing doors. In this paper we will not discuss the filtering techniques that can be applied to the received audio signals in order to reduce the influence of noise on the audio source locator.

The *idealized environment* of the audio source locator is defined in this paper for ease of reasoning. In the ideal case an audio source is considered to be a single point in space. Further, for the ideal case it is assumed that there is exactly one audio source emitting sound (ignoring the problem of interfering speakers and background noise). Finally, sound is assumed to propagate from source to receivers with equal speed and intensity for all directions and received by the microphones with equal intensity[2].

2.3 Requirements

Summarizing, we can state that the audio source location problem reduces to estimating time-delays between pairs of received audio signals. The requirements for the accuracy of the source location estimation originate from the digital TV-director application as described in (Mullender 1994). A summary of these minimal requirements is:

- Usage of a minimal amount of resources: preferably using standard workstations only and a minimal number of microphones;
- Accuracy: the audio source must be located within an angle $\alpha \pm 5$ degrees;
- High estimation rate: four times per second;
- Low estimation delay;
- Robustness to background noise.

3 HUMAN SPEECH

The design of the algorithms presented in this article is based on investigation of human speech properties. This section describes these properties and the implications for the designs based on these properties.

3.1 Properties of human speech

Basically, speech can be subdivided into two types; *voiced* speech and *unvoiced* speech. In voiced speech, the main speech energy source is the vibration of the vocal cords. The frequency of this vibration is the same as the fundamental frequency of the speech signal and determines the pitch of the voiced phonation. In addition to the fundamental frequency component, its harmonics (with intensity decaying at 12 dB/octave approximately) are generated near the vocal cords.

1. On average the speed of sound through air is 343 m/s. It varies for instance with air temperature, pressure and humidity.

2. In pratice this assumption does not hold (see sections 5.2 and 6.2)

The acoustic tube of the mouth, nose and the other articulatory organs, called the vocal tract, act as a resonant filter on these frequencies. Changing the shape of the vocal tract is called articulation. In addition to vocal cord vibrations, a significantly smaller amount of speech energy originates from air-turbulence in the vocal tract. In voiceless (unvoiced) speech the only speech energy source is air turbulence.

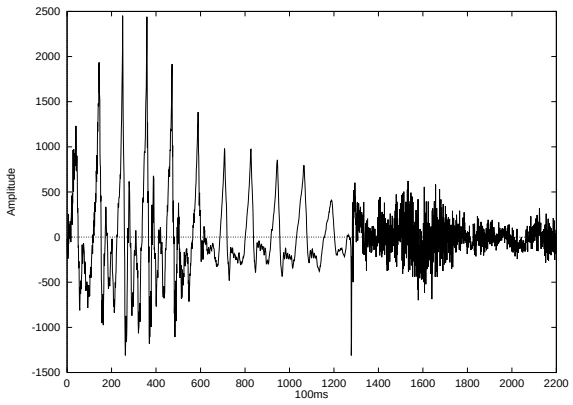

Figure 3: Sample of Voiced (left part) and Unvoiced (right part) speech of a sample interval of 100 msec (2200 samples)

From Figure 3 we can see the different appearances of voiced and unvoiced speech. It is shown in this figure that voiced speech is more or less periodic with a period equal to the fundamental wavelength, while unvoiced speech is non-periodic and has a rather noise-like nature. Due to the presence of the powerful fundamental frequency and its harmonics in voiced speech, voiced speech has significantly more power than unvoiced speech.

The *fundamental frequency* is the lowest (significant) frequency present in voiced speech. Studies have shown that the fundamental frequency varies continuously and slowly in time for conversational speech. The average fundamental frequency for individual male speakers is about 250 Hz, and for female speakers about 300 Hz.
The time ratio for voiced, unvoiced and silence intervals in speech is roughly 60%/25%/15% for normal conversational speech, and unvoiced speech intervals are short in duration most of the time (Cook 1991; Saito and Nakata 1985).

3.2 Implications for audio location

In order to show the suitability of both types of speech for location (i.e. time-delay estimation) purposes, the auto-correlation function (AC) for both speech fragments is depicted in Figure 4. The autocorrelation function of a signal is the cross-correlation function of a signal with itself, and is a good measure for the equality of a signal and its time-shifted variants. The maximum of the autocorrelation function will therefore be located at shift zero, as can be seen in the two figures.

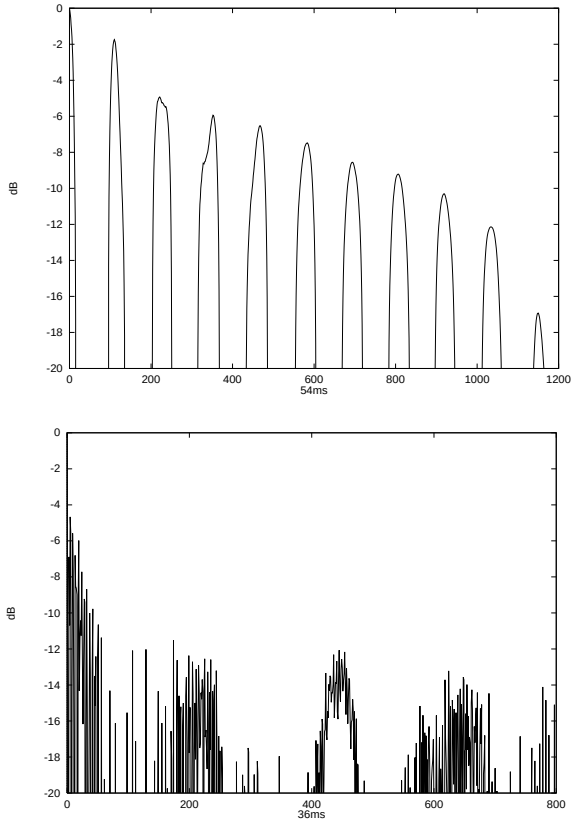

Figure 4: Autocorrelation of Voiced Speech part (top) and Unvoiced Speech part (bottom).

For voiced speech the autocorrelation function has a clear maximum peak at shift zero. Consecutive peaks however can be found at multiples of the fundamental wavelength, caused by the high-intensity fundamental frequency and its harmonics. These secondary peaks are unfortunately hardly less intense than the main peak located at shift zero. In practical, non-ideal situations these secondary peaks can easily become more intense than the 'main' peak, causing the most intense peak in the correlation not to be associated with the time-delay we are looking for. Straightforward cross-correlation of voiced speech signals is therefore only applicable on a limited range of possible time shifts. Within this range of approximate size equal to the fundamental wavelength, time-delay estimation can be performed quite accurately though.

For unvoiced speech the autocorrelation function gains its maximum value at shift zero also. However, the autocorrelation function decays less fast around this maximum value as the voiced speech autocorrelation. Furthermore, no intense secondary peaks appear when correlating unvoiced speech, due to the absence of a fundamental frequency and its harmonics. This implies that time-delay estimation using unvoiced speech is potentially less accurate than estimation using voiced speech, but that the estimation range is not restricted by the nature of unvoiced speech. Unfortunately, the time ratio for voiced, unvoiced and

silence intervals in speech is roughly 60%/25%/15%, and unvoiced speech intervals are short in duration most of the time. This implies that unvoiced speech intervals are also not very suitable for wide-range time-delay estimation.

4 RANGE AMBIGUITY PROBLEM

The required time-delay range corresponding to a certain setting is determined by the distance between the microphone pairs d_m. The maximum and minimum time-delay possible for a certain setup are linear dependent to this distance, according:

$$-\frac{d_m}{c} \leq \theta_{ij} \leq \frac{d_m}{c} \qquad (1)$$

with c denoting the propagation speed of sound through air. These bounds to the size of the expected time-delays can now be used to determine the frequencies in the received signal that will contribute to ambiguity in the correlation results. A frequency can contribute to the ambiguity when its wavelength is shorter than the range of the expected time-delay. The highest frequency that is still unambiguous with regards to time-delay estimation for a certain microphone distance is then defined by:

$$f_{high} = \frac{c}{2d_m} \qquad (2)$$

Any frequency in the received signal higher than f_{high} will contribute to the ambiguity of the time-delay estimation. However, a secondary peak in the correlation sum will only be caused by a frequency when its power and the power of its harmonics is relatively high. The fundamental frequency is such a frequency for voiced speech as has been shown in the previous section.

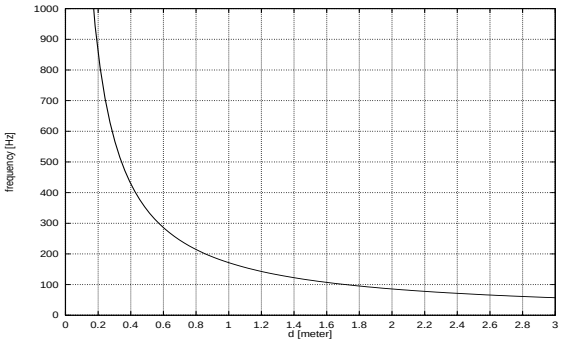

Figure 5: Highest unambiguous frequency as function of microphone distance.

From Figure 5 we derive that f_{high} decreases quite fast when the microphone distance increases. Already at a microphone distance of a few decimetres, f_{high} becomes lower than the fundamental frequency of a lot of speakers. Therefore, only a microphone spacing of a few centimetres (*closely spaced* i.e. < 20 cm), will avoid the powerful periodic components in (voiced) speech from causing ambiguity in the correlation results. However, placing a pair of microphones close together increases the eccentricity of the time-delay hyperbolae (see Figure 2) which has a negative effect on the accuracy of the coordinate calculations.

In the next sections we will present three audio source location algorithms. The first is based on straight forward *cross-correlation* of speech with widely spaced microphone-pair (between 1 - 2 m). The second is based on *two stage* cross-correlation of speech where the periodic component has been filtered out initially. In this approach the distance between the microphones is also about 2 meters. Finally an algorithm is presented that used the voiced part of speech only with closely spaced microphones, using a *high-speed correlator* for the required accuracy.

5 CROSS-CORRELATION

5.1 Introduction

First we will consider a technique for time-delay estimation assuming the idealized environment described in a previous section. For the idealized environment we may assume that the received signals are identical except for a certain time-delay between them.
One technique to measure the time-delay of a signal and a time-shifted version of it, is by cross-correlating these signals (Sijben 1993). The cross-correlation of two signals is a measure for how well these signals match for different shifts of one of them. Best fit of the two signals is found at that shift where the cross-correlation gains its maximum. Cross-correlation is performed on two received signals $r_i(t)$ and $r_j(t)$ according:

$$CC_{ri,rj}(\tau) = \sum_{n=0}^{N-1} r_i(nT)\, r_j(nT+\tau) \qquad (3)$$
$$\tau = \ldots,-T,0,T,\ldots$$

In this equation τ represents the time-shift, r_i and r_j are the received signals, T is the sampling period and N is the number of samples used for the correlation. In order for cross-correlation to be useful for time-delay estimation purposes, the assumption needs to be made that when the cross-correlation function of two shifted versions of the same original signal exhibits an absolute maximum, then its corresponding shift-value denotes the actual time-delay between these two versions. This is not generally valid for all possible signals when Equation (3) is used. This equation might for example exhibit higher values at other (higher) time-shifts, when the power of the signal increases with time. This may easily lead to an incorrect time-delay estimation. The applicability of the cross-correlation function as a time-delay estimator therefore depends on the properties of the signals being correlated. A major disadvantage of this method is the com-

plexity of the cross-correlation calculation: O(N^2). This requires considerable computing power for large N.

5.2 Experiments with the TV-director

The algorithm was used in a first version of the digital TV-director, called 'Federico'. The experiments were performed using sampling rates from 8kHz to 22kHz.

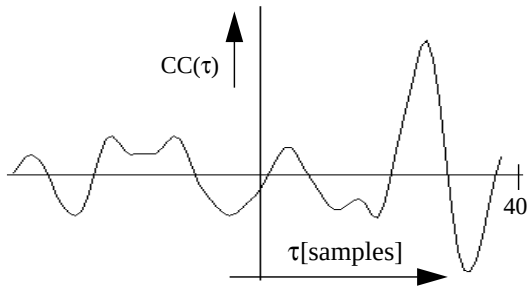

Figure 6: The cross-correlation of two typical microphone signals.

Figure 6 shows the cross correlation of the two typical signals. The peak in the correlation indicates the estimated time-delay.
The range ambiguity problem, as described before, produces correlations from which the main peak is difficult to distinguish. Figure 7 shows a correlation of a

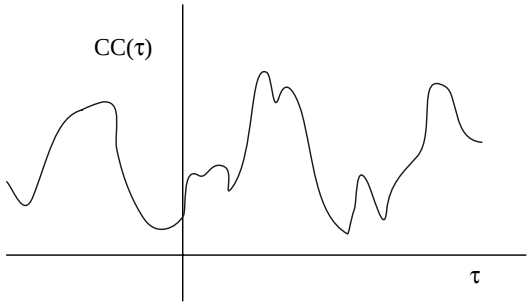

Figure 7: Correlation of a signal in which the highest top is only slightly higher than its neighbours

such an input signal. In this setting the time-dealy may not be more than 40 samples.

Several methods were tried to avoid the ambiguity problem. Post-processing filters were used to separate the main correlation peak from secondary and tertiary peaks. Only peaks with a value larger than a certain value are accepted. Statistical filters, called 'sceptics', were used to filter out incidental erroneous values that suggested locations that were far from previous measured locations.
With these filters adequate results for speakers that spoke loud and clear were obtained. The accuracy of the algorithm is sufficient but the robustness was insufficient.

6 2-STAGE CROSS-CORRELATION

As shown in the previous sections, straightforward cross-correlation is an accurate and efficient time-delay estimation technique for the *idealized* environment. For most practical purposes, however, its lack of robustness (when using widely-spaced microphone pairs) and/or accuracy (when using closely-spaced microphone pairs) limits the application and environment in which it can be used. In order to meet all our requirements as stated in section 2.3, we need to deal with the range-ambiguity problem. Several approaches can be followed to reduce or avoid ambiguity in the correlation results. Some approaches require modifications of the basic minimal setup (e.g. by adding physical resources). A number of approaches are described and discussed in (Dillema 1994).

In this section, we describe a general approach not requiring any adjustments to the basic setting. In section 7 an approach is described that makes a few adjustments to the basic setting in order to meet its requirements by adding additional hardware.

6.1 Separating periodic and non-periodic components

Dealing with the range ambiguity in case of widely-spaced microphone pairs requires the application of additional techniques and/or more subtle application of the correlation technique. Section 4 described the main cause of ambiguity in the correlation results, viz. the high-energy periodic component of voiced speech. Although unvoiced speech and the noise-like component contained in voiced speech span the high-frequency range and therefore are contributors to ambiguity in the results, their random nature prevents these components to cause powerful secondary (ambiguous) peaks in the correlation function. Filtering the periodic component from voiced speech and using the residual for time-delay estimation (using cross-correlation), will therefore reduce the ambiguity in the results considerably. Using the assumptions on the idealized environment, this can be formalized as follows:
Let us assume that we can perfectly split the received signals into a periodic component and a non-periodic component signal, then:

$$r_i(nT) = p(nT) + np(nT), \text{ and}$$
$$r_j(nT) = r_i(nT + \theta) = p(nT + \theta) + np(nT + \theta)$$

Then according (Dillema 1994):

$$CC_{ri,rj}(\tau) = AC_p(\tau + \theta) + AC_{np}(\tau + \theta) + CC_{p,np}(\tau + \theta) + CC_{p,np}(\tau - \theta)$$

in which AC_x is the autocorrelation function of x and $CC_{x,y}$ the cross-correlation function of x and y (see 3.2) When we assume that the periodic and non-periodic component form disjoint sets in the frequency spectrum, the last two terms in the above equation are zero. Then:

$$CC_{ri,rj}(\tau) = AC_p(\tau + \theta) + AC_{np}(\tau + \theta), \text{ or}$$

$$CC_{ri,rj}(\tau) = CC_{pi,pj}(\tau) + CC_{npi,npj}(\tau)$$

In other words, when separating the periodic and non-periodic components of the received signals we can use either cross-correlation of the periodic components, or cross-correlation of the non-periodic components or both.

Not using the periodic component of voiced speech implies, however, not using the high-power components of the speech signal. This means that only a small portion of the dynamic range of the samples is used. In addition, the signal-to-noise ratio will be much lower, making the calculations much more sensitive to increasing noise-levels. For example, the algorithm is especially more vulnerable to murmuring and whispering people. At the same time, peak-picking becomes more difficult and less accurate, due to the small peak-width of the correlation maximum for noise-like (unvoiced) speech as described in section 3. All these factors will have a negative effect on the accuracy of the time-delay estimations. Summarizing, we can state there is a trade-off between estimation accuracy and ambiguity when a periodic-component filter is first used to reduce ambiguity.

A *two-stage algorithm* is useful to bridge the trade-off between accuracy and ambiguity of the time-delay estimation results. The first stage of this algorithm cross-correlates the filtered, non-periodic audio signals yielding an unambiguous, but not very accurate time-delay estimate. The second stage then uses this initial estimate to resolve the range-ambiguity of the straightforward cross-correlation (i.e. without filtering of the periodic component). The initial estimate is used as a range limiter for the final time-delay estimation, yielding an estimate that is without ambiguity but at a high accuracy.

6.2 Implementation

Implementing a filter that can separate the periodic (fundamental frequency and its harmonics) and non-periodic component of the received signals is not trivial. Two different approaches can be followed in designing such a filter. The first approach tries to estimate the fundamental frequency based on the received audio signals, and uses this estimation to separate the fundamental frequency and its harmonics from the rest of the signal. The second approach uses a priori known properties or heuristics on speech to build a filter that roughly separates the periodic and non-periodic component. Our implementation is based on the latter approach, yielding a (computational) very simple filter. While in general most energy of the periodic component is contained in the lower frequency range and most energy of the non-periodic component is in the higher-frequency range, a simple high-pass filter is used to remove most of the periodic component from the signal. In Figure 8 we illustrate the effect of a high-pass filter with cut-off frequency at 2.5 kHz, where the intensity difference between the most intense secondary peak and the main peak is increased from merely 1.2 dB to 5.4 dB.

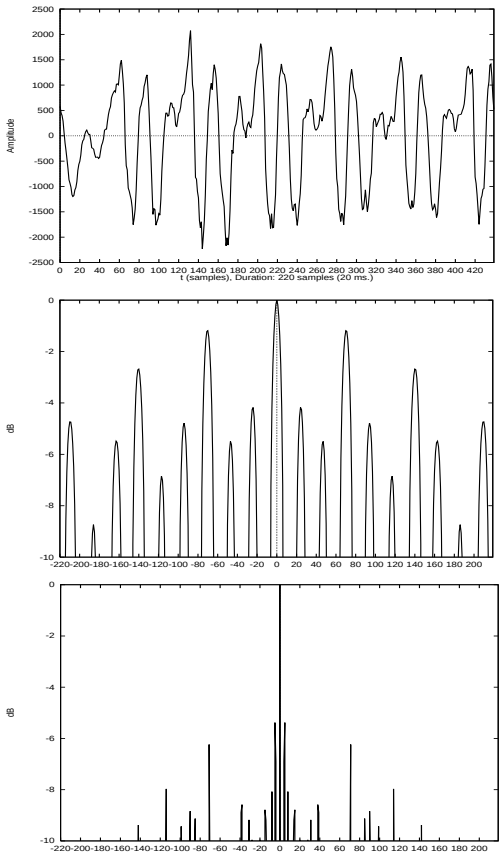

Figure 8: top: Voiced Speech Fragment, middle: Its frequency domain autocorrelation, bottom: Frequency domain autocorrelation when simple filter is applied first

We have implemented a two-stage time-delay estimator as described in this section. It performs cross-correlation and filtering in the frequency domain, so that these operations have a computational complexity of O(*N*), where *N* is the number of samples per block used for each estimation. The Fast Fourier Transform is used to transform from the time domain to the frequency domain, bringing the computational complexity of the time-delay estimator to O(*N*log(*N*)).

The idealized environment assumes that the signals received by different microphones are identical except for a certain time shift. In practice this assumption is violated for a number of reasons, but mainly because sound intensity decreases inversely proportional to the square of the distance from the source. The assumption can be validated in practice by power normalizing the sampled signals or by using an automatic gain controller before sampling. We chose for the latter in order to make better use of the dynamic range of the sampling equipment.

The speech interval used for each estimation needs to

be only a few periods of the fundamental frequency of the speaker (between 10 and 20 msec.) resulting in a high maximum estimation rate (between 50 and 100 estimations per second). The actual rate depends of course also on the available computational power, but our current implementation indicates that the maximum rate can be achieved with the computational power of standard workstation or desktop PC. The accuracy of the time-delay estimates has not been thoroughly analysed yet. Preliminary test results, however, indicate that the accuracy of the time-delay estimator meets the requirements of the TV-director application in reasonable and realistic environments. These results also indicate that the robustness of this algorithm is better than our previous approach.

7 HIGH-SPEED SIGNAL CORRELATION ALGORITHM

7.1 Introduction

In this algorithm the voiced part of speech is used to correlate the signals from two microphones. Only the fundamental frequency of speech is used. As apposed to the previous approaches the microphones for this algorithm are placed close together, in the prototype about 12 cm. This has several advantages:

- the signals received from both microphones are strongly correlated and have almost the same shape, i.e. the assumptions of the idealized environment are valid;
- the range ambiguity problem (see section 4) is not present[1];
- because the distance between the microphones is small, the microphones can be placed on the pan/tilt device. This means that the microphones can be directed towards the location of the speaker. When the source of speech lies in the middle of the microphone pair the accuracy is maximal (see Figure 11).

However, to achieve the required accuracy we need high speed sampling.

7.2 Design

Figure 9 gives the block diagram of the correlator. The amplified signals from the microphones are passed to a second order low pass filter with a cut-off frequency of 500 Hz (which is less than f_{high}). These signals are compared with a threshold *Vref* to eliminate low amplitude signals (e.g. noise during silence intervals).[2] A micro-controller receives the resulting signals (S_{M1} and S_{M2}) from which it calculates the time-

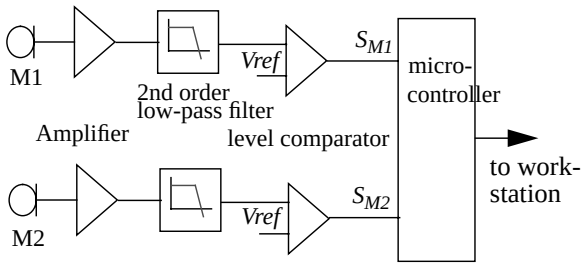

Figure 9: Block diagram of the high-speed correlator

delay θ_i (see Figure 10). The time-delay θ_i is only valid when $0 < \theta_i < (d_m.29\mu s^3)$. Non-valid values are discarded.

The micro-controller starts sampling when the first pulse on signal S_{M1} or S_{M2} arrives. Then, during 200 ms time-delays θ_i are calculated. On an average voiced speech interval this will give about 80 valid time-delays θ_i sufficient to calculate a useful time-delay distribution. From this distribution the estimated

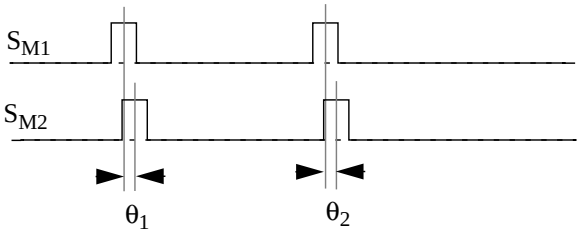

Figure 10: Time-delay between M1 and M2

time-delay θ, the variance and the confidence of the time-delay can be calculated in various ways. Further study is necessary to select the most suitable method. The time-complexity of the algorithm is *O(n)* and can easily be computed on a low cost micro-controller (*n* is the number of time-delays θ_i).

Every 250 msec. the workstation can get a new time-delay estimation including variance and confidence level. A prototype of this design has been built and has been successfully tested giving the expected results. Currently we are designing a prototype with four microphones to increase accuracy and speed[4].

7.3 Sample frequency

To achieve the required accuracy we need high speed sampling. When *m1* is the distance between the audio source and microphone M1, and *m2* is the distance between the audio source and microphone M2, then Figure 11 gives the difference δ between *m1* and *m2* versus angle α of the audio source (see Figure 1). For the estimation of the required sample frequency

1. The highest frequency that is still unambiguous (f_{high}) in this setup equals to 1.4 kHz.
2. The level of *Vref* can be adjusted to eliminate the background noise.
3. The speed of sound through air is approximately 343 m/s, or in other words 1 cm per 29 µs.
4. The speed can be improved because more microphones give more time-delay estimations per second.

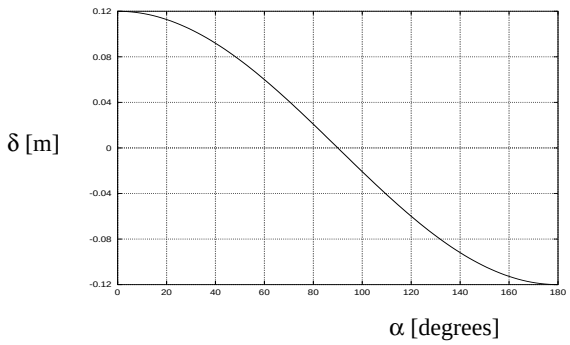

Figure 11: δ = (m1-m2) versus audio source angle α (d_m=12 cm, d=300 cm)

we assume a linear relation between δ and the angle α. To aim a camera with an accuracy of about 5 degrees, we need to be able to distinguish 180/5=36 sectors (assuming the audio angle is between 0 and 180 degrees). Given that the distance between the microphones d_m = 12 cm, δ changes from 12 to -12 cm when α changes from 0 to 180 degrees. So this gives (12+12)/36 = 0.66 cm per sector; which corresponds to a time-delay of 0.66 * 29 = 19 μs per sector. So, when we also incorporate the quantisation error, we need a sampling rate of at least 100kHz (10 μs) to discriminate between the sectors.

8 CONCLUSION

Time-delay estimation is an useful technique to estimate an audio source location. This has already been shown in related research, but their design goals and requirements were different from ours. The approaches taken in this paper faces audio source location from different points of view than most related work in this area. By limiting the scope of the system purely to location of human speakers, an accurate audio source locator is shown to be feasible, utilizing the characteristic properties of speech signals, like differences between voiced and unvoiced speech.

The first algorithm uses straightforward cross-correlation. The range ambiguity problem produces correlations from which the main peak is difficult to distinguish. Several methods and filters are needed to give satisfactory location results. To acquire unambiguous robust results severe restrictions on the setting of the audio source locator can be opposed. With the second algorithm we have shown that less restrictive techniques can be applied to extend the range of the time-delay estimation, making a location system feasible needing few resources.
A *two-stage algorithm* is useful to bridge the trade-off between accuracy and ambiguity of the time-delay estimation results. The first stage cross-correlates the filtered, non-periodic audio signals yielding an unambiguous, but not very accurate time-delay estimate. The second stage then uses this initial estimate to resolve the range-ambiguity of the straightforward cross-correlation. The initial estimate is used as a range limiter for the final time-delay estimation, yielding an estimate that is without ambiguity but at a high accuracy.

Another approach was taken with the high speed correlator. With a high sampling rate and pre-filtering the required accuracy was reached, even with the microphones close together. The charm of the high-speed correlator approach lies in its simplicity, resulting in a low-cost design with little overhead for a workstation. It is able to provide a time-delay estimation together with variance and confidence level every 250 msec.

9 BIBLIOGRAPHY

Brandstein M.S., Adcock J.E., Silverman H.F.: "A closed-form method for finding source locations from microphone-array time-delay estimates", *Proceedings ICASSP-95*, pp 3019-3022, IEEE, 1995.

Cook, P.R., "Identification of control parameters in an Articulatory vocal tract model, with applications to the synthesis of singing", department of EE, Stanford University, september 1991

Dillema, F.W., "Audio Source Location", Masters Thesis, University of Twente, The Netherlands, March 1994.

Omologo M., Svaizer P.: "Acoustic source location in noisy and reverberant environment using CSP analysis", *Proceedings ICASSP-96*, May 7-10, 1996.

Mullender, S.J., "Specification of the Digital TV Director", Pegasus Paper, University of Twente, The Netherlands, September, 1994. (see also: http://www.pegasus.esprit.ec.org/papers/pegpapers.html)

Saito, S. and Nakata, K., "Fundamentals of Speech Signal Processing", Academic Press, Tokyo, 1985

Sijben, P., "Audio Source Location", Masters Thesis, University of Twente, The Netherlands, January 1993.

Brandstein, M.S. and Silverman, H.F., "A New Time--Delay Estimator for Finding Source Locations using a Microphone Array", Technical Report LEMS-116, Division of Engineering, Brown University, March 1993.