CLASSIFICATION OF TIME DELAY ESTIMATES FOR ROBUST SPEAKER LOCALIZATION

N. Strobel and R. Rabenstein

University of Erlangen-Nürnberg, Telecommunications Laboratory
Cauerstr. 7, 91058 Erlangen, Germany
strobel@nt.e-technik.uni-erlangen.de

ABSTRACT

This paper proposes a solution to the problem of robust speaker localization under adverse acoustic conditions. The approach is based on the classification of time delay estimates. Two classification techniques are investigated in detail: maximum likelihood (ML) classification and classification based on histogram comparison. Their performance under adverse acoustic conditions is compared to outcomes obtained with the traditional approach which uses time delay estimates directly to infer speaker positions. Experiments indicate that the ML classification method provides little improvement over the traditional method. On the other hand, using histogram classification, we can improve the probability of correct speaker localization by more than 60% compared to either the traditional approach or the ML classification technique.

1. INTRODUCTION

Different solutions have been suggested to tackle the problem of passive speaker localization. They normally rely on the acquisition of time-delayed replicas of a source signal at spatially distributed sensors. The generalized cross-correlation technique is commonly applied to obtain time differences of arrival (TDOAs) between multiple sensor signals [1]. TDOA estimates are subsequently used as parameters specifying the source position. Chan and Ho proposed a technique based on intersections of hyperbolic curves [2]. Brandstein et al. used the linear intersection (LI) estimation method [3]. Both techniques yield closed-form solutions. In [4], Wang and Chu presented the voice source localization system used in the PictureTel automatic camera pointing system. They average signal onsets and apply the phase correlation technique which has been found to perform adequately for acoustic event localization in real environments. However, for adverse noise and reverberation conditions, Omologa and Svaizer showed experimentally that there are circumstances under which the phase correlation techniques no longer yields reliable results [5].

In this paper we present an approach to speaker localization especially designed for difficult acoustic conditions such as severe multipath distortion (reverberations), acoustic obstructions, and background noise. Similar to conventional methods, TDOAs are still taken as inputs for subsequent processing steps. Position estimates are, however, no longer based on geometrical considerations under ideal conditions. Instead they follow by classification of time delay estimates obtained under multipath propagation and in the presence of background noise.

This paper is organized as follows: First, we briefly outline time delay estimation. Second, we describe two techniques for source localization based on the classification of time delay estimates. Third, we present the performance of our approach based on experiments in an anechoic chamber. Finally, a discussion of the results is provided and some conclusions are drawn.

2. TIME DELAY ESTIMATION

Consider two received sensor signals $y_i(t)$ and $y_j(t)$ originating from a common source as shown in Fig. 1. The distance from the i-th microphone to the source is specified as $\mathrm{d_i}$, and the source is positioned $\mathrm{d_j}$ away from the j-th microphone. The distance between the two sensors is denoted as d. Assuming far-field conditions and that d is smaller than half of the wave length, the direction of arrival can be unambiguously computed from

$$\cos(\theta) = \frac{\tau_{ij} \cdot c}{\mathrm{d}}. \tag{1}$$

In Eq. (1), τ_{ij} is the time delay between the microphones i and j, while c denotes the speed of sound. For nondispersive wave propagation we get $\tau_{ij} = \frac{\mathrm{d_i} - \mathrm{d_j}}{c}$.

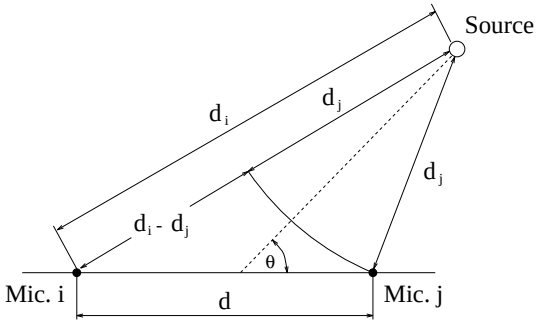

Figure 1: Configuration used to explain the time difference of arrival (TDOA) between signals recorded at the i-th and j-th microphone.

One of the most popular techniques for TDOA estimation is the generalized cross-correlation technique. Although optimal for single path propagation of Gaussian signals contaminated by uncorrelated white noise, its performance can deteriorate significantly when echos are present. In this case, the signal at the i-th microphone recorded when there are M echos and ambient noise can be

written as

$$y_i(kT_s) = s_0(kT_s - \tau_{0i}) + \sum_{m=1}^{M} \alpha_m \, s_0(kT_s - \tau_{mi}) + n_i(kT_s), \; i = 1, \ldots, Q. \quad (2)$$

The waveform $s_0(kT_s)$ represents the desired speech signal. It is received and sampled by Q omnidirectional microphones with suitably chosen sampling period T_s. Further, α_m is the unknown reflection coefficient of the m-th reverberation, and $n_i(kT_s)$ is ambient noise at the i-th microphone. The time delays τ_{0i} and τ_{mi} are associated with the propagation time from the talker and the m-th reverberation to the i-th microphone. Figure 1, e.g., shows that $\tau_{0i} = \mathrm{d_i}/c$. Following common notation, the discrete-time signals $y_i(kT_s)$ and $y_j(kT_s)$ are further on denoted as $y_i[k]$ and $y_j[k]$, respectively.

Due to the nonstationary behavior of speech, a short-time signal analysis is needed. To that end, the total record length of the i-th and j-th sensor outputs is divided into K disjoint segments (frames) of (even) length L. They are indexed using κ as the frame number, $\kappa = 1 \ldots, K$. The frame number is related to the center sample of the κ-th time segment according to $k = (2\,\kappa - 1)\,\frac{L}{2}$.

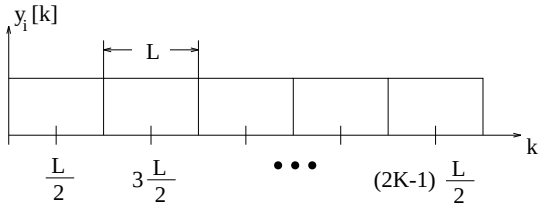

Figure 2: For short-time TDOA estimation, the input signals are divided into K time segments (frames). Each frame contains L samples. .

Having introduced signal frames and their numbering, we can formulate the TDOA estimation procedure as follows:

First, obtain estimates of the short-time power spectra $\hat{G}_i(\kappa, \omega)$ and $\hat{G}_j(\kappa, \omega)$ by evaluating

$$\hat{G}_i(\kappa, \omega) = \sum_m y_i[\kappa + m]\, w[m]\, e^{-j\omega m}. \quad (3)$$

In Eq. (3), $w[.]$ is a (rectangular) window sequence with (even) length L, symmetric around the frame center.

Next, compute the short-time estimate of the generalized cross correlation function between the i-th and j-th sensor signal from

$$\hat{R}_{ij}[\kappa, l] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \psi_{ij}(\kappa,\, \omega)\, \hat{G}_{ij}(\kappa,\, \omega)\, e^{j\omega l}\, d\omega. \quad (4)$$

In Eq. (4), $\psi_{ij}(\kappa,\, \omega)$ is the frequency weighting filter for the κ-th frame. The estimate of the cross-power spectrum of $y_i[k]$ and $y_j[k]$ is given by

$$\hat{G}_{ij}(\kappa,\, \omega) = \hat{G}_i(\kappa,\, \omega)\, \hat{G}_j^*(\kappa,\, \omega). \quad (5)$$

Similar to [4] and [5], we restrict ourselves to the phase correlation method with

$$\psi_{ij}(\kappa, \omega) = \frac{1}{|\hat{G}_{ij}(\kappa,\, \omega)|}. \quad (6)$$

Finally, the estimated delay between $y_i[k]$ and $y_j[k]$ for the κ-th signal frame, $\hat{l}_{ij}[\kappa]$, is the sample lag maximizing Eq. (4):

$$\hat{l}_{ij}[\kappa] = \arg\max_l \{\hat{R}_{ij}[\kappa, l]\}. \quad (7)$$

The associated time difference of arrival (in seconds) for the κ-th segment is

$$\hat{\tau}_{ij}[\kappa] = \hat{l}_{ij}[\kappa] \cdot T_s. \quad (8)$$

Substituting $\hat{\tau}_{ij}[\kappa]$ into Eq. (1), we obtain an estimate of the speaker position for each frame. Unfortunately, such a direct approach is likely to fail under adverse acoustic conditions, since accurate TDOA estimates are necessary for Eq. (1) to hold. More robust position estimates can, however, still be obtained if time delay estimates are classified as explained next.

3. TIME DELAY CLASSIFICATION

When multipath wave propagation is present, many replicas of each wave front reach the sensors. This makes time delay estimation more difficult and in some cases impossible. However, when wave propagation takes place within a known environment, conditional probability density functions (pdfs) can be used to improve source localization. They can either be obtained using the image model [6], or they can be learned directly from measurements during a training period.

Once we know the conditional pdfs of time delay estimates given all representative speaker positions, we can improve source localization by time delay classification. To this end, we consider a maximum likelihood (ML) approach first. Afterwards, we discuss classification by histogram comparison. For simplicity, the algorithms are explained for plane waves impinging on an array consisting of two microphones as shown in Fig. 1. An extension of the basic ideas to more sophisticated arrays is conceptually straightforward.

Each conditional probability density function describes a distribution of time delay estimates for a given source position in a known acoustic environment. To estimate the conditional pdfs, we first divide the whole horizon into P angular sectors. Each sector is represented by its center angle $\theta^{(p)}$. The superscript p indexes all P angular regions, i.e., $p = 1, \ldots, P$.

The (integer) time delay between the i-th and the j-th microphone associated with the p-th angular region, $l_{ij}^{(p)}$, follows from

$$l_{ij}^{(p)} = \left[\frac{\cos(\theta^{(p)}) \cdot d}{c} \cdot \frac{1}{T_s}\right]. \quad (9)$$

The symbol [.] denotes rounding to the next nearest integer. For simplicity, we set up adjacent angular regions, $\theta^{(p)}$ and $\theta^{(p+1)}$, such that associated time delays, $l_{ij}^{(p)}$ and $l_{ij}^{(p+1)}$, differed by a constant number of samples, specified as Δ. That is,

$$|l_{ij}^{(p)} - l_{ij}^{(p+1)}| = \Delta = \text{const}\ \forall p. \quad (10)$$

Since the continuous conditional pdfs $p(l_{ij} \,|\, \theta^{(p)})$ are not available, histograms $H(l_{ij} \,|\, \theta^{(p)})$ are used instead. They contain P bins with (identical) bin width Δ. The bins are centered around the time delays $l_{ij}^{(m)},\ m = 1, \ldots, P$. The superscript, m, in $l_{ij}^{(m)}$ refers to the m-th histogram bin representing the m-th angular region. There are as many histogram bins as there are angular regions. To populate the bins, we position a source under some incident angle $\theta^{(p)}$. Then we use Eq. (7) to estimate K time delays,

$\hat{l}_{ij}[\kappa]$, $\kappa = 1, \ldots, K >> 1$. Each time delay estimate $\hat{l}_{ij}[\kappa]$ is added to the appropriate histogram bin by effectively quantizing it to the closest $l_{ij}^{(m)}$ and incrementing an associated counter. The number of discrete TDOA estimates finally found in the m-th bin centered around $l_{ij}^{(m)}$ is called $n_m^{(p)}$. The subscript m in $n_m^{(p)}$ recalls the associated (m-th) histogram bin. The superscript p is reminiscent of the incident angle $\theta^{(p)}$ the histogram has been estimated (trained) for.

In the limit, the conditional histogram $H(l_{ij}|\,\theta^{(p)})$ for a given incident angle $\theta^{(p)}$ can be viewed as an ensemble of estimated discrete probabilities arranged over all histogram bins. An example is shown in Fig. 3.

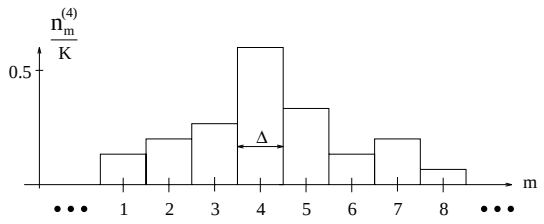

Figure 3: Idealized histogram obtained for a source under incident angle $\theta^{(4)}$. The histogram shows that most of the estimated time delays, $l_{ij}^{(m)}$, fall into the appropriate histogram bin with $m = 4$. This is the case when there are no significant echos.

3.1. Maximum Likelihood Classification

The maximum likelihood (ML) classifier takes the time delay estimate for the κ-th frame, $\hat{l}_{ij}[\kappa]$, as classifier input and returns an instant estimate of the associated angular region, $\hat{\theta}_{ML}[\kappa]$, which is most likely for $\hat{l}_{ij}[\kappa]$. More mathematically,

$$\hat{\theta}_{ML}[\kappa] = \arg\max_{\theta^{(p)}}\{H(\hat{l}_{ij}[\kappa]|\,\theta^{(p)})\},\ p = 1, \ldots, P. \qquad (11)$$

Recall that the angular regions are represented by their center angles $\theta^{(p)}$, $p = 1, \ldots, P$. As a result, $\hat{\theta}_{ML}[\kappa]$ can only assume P different values.

Using histograms, we implement ML classification by means of a contingency table. That is, given a particular time delay, we look up the most likely angular region. The tables are based on the histograms $H(l_{ij}|\,\theta^{(p)})$ obtained during the training period.

The ML classification procedure improves source localization when the direct path from the source to the microphone array is obstructed. In the presence of strong echos, a much more successful strategy is to postpone source localization until many frames have been evaluated. Such a technique is explained next.

3.2. Classification by Histogram Comparison

For objects slowly moving compared to the sampling rate, many frames may be used to arrive at an estimate for a source position. We propose classification by histogram comparison as a practical implementation of this idea. The major difference to the ML method is that this method no longer relies on the classification of a single TDOA estimate. Instead, it takes an ensemble of many estimates and arranges them as a histogram which is finally classified to obtain an estimate of the unknown angular region of arrival.

Let the measurement histogram be associated with the unknown angle of arrival $\theta^{(q)}$. In a first step, N signal segments (measurement frames) are recorded each consisting of L samples. Next, we estimate N associated time delays $\hat{l}_{ij}[\kappa]$, $\kappa = 1, \ldots, N$. Their statistical distribution yields a measurement histogram referred to as $\hat{H}(l_{ij}|\,\theta^{(q)})$.

Once the measurement histogram $\hat{H}(l_{ij}|\,\theta^{(q)})$ has been computed, we find the unknown source direction $\theta^{(q)}$ by comparing $\hat{H}(l_{ij}|\,\theta^{(q)})$ to the training histograms $H(l_{ij}|\,\theta^{(p)})$, $p = 1, \ldots, P$, introduced earlier. Simply choosing the best fit, we obtain a very robust estimate for the unknown angle of incidence $\theta^{(q)}$.

The chi-squared metric is used as a distance measure between the measurement histogram $\hat{H}(l_{ij}|\,\theta^{(q)})$ and the p-th training histogram $H(l_{ij}|\,\theta^{(p)})$ associated with $\theta^{(p)}$. For the χ^2 comparison of the measurement histogram to the p-th training histogram, we get

$$\chi^2(\theta^{(q)}|\,\theta^{(p)}) = \sum_{m=1}^{P} \frac{(N_m^{(q)} - \hat{n}_m^{(p)})^2}{\hat{n}_m^{(p)}}. \qquad (12)$$

In Eq. (12), $N_m^{(q)}$ denotes the number of time delay estimates in the m-th bin of the measurement histogram $\hat{H}(l_{ij}|\,\theta^{(q)})$. The expected number of time delay estimates in the m-th bin based on the p-th training histogram, $\hat{n}_m^{(p)}$, follows from

$$\hat{n}_m^{(p)} = \frac{n_m^{(p)}}{K} \cdot N. \qquad (13)$$

Equation (13) merely takes $n_m^{(p)}$ recorded over K training frames and adjusts it to N, the number of measurement frames ($N < K$).

Finally, the χ^2-estimate of the incident angle, $\hat{\theta}_{\chi^2}$, is the angle whose value for $\chi^2(\theta^{(q)}|\,\theta^{(p)})$ is smallest, i.e.:

$$\hat{\theta}_{\chi^2} = \arg\min_{\theta^{(p)}}\{\chi^2(\theta^{(q)}|\,\theta^{(p)})\},\ p = 1, \ldots, P. \qquad (14)$$

Since there are only P different training histograms each characterizing a particular angular region, we only get P different values for $\hat{\theta}_{\chi^2}$.

4. EXPERIMENTS AND RESULTS

The experiments took place in an anechoic chamber. The array consisted of two microphones spaced approximately $1.3\,\mathrm{m}$ apart. To introduce echos, the walls to the left and right of the speaker were modified by adding material reflecting any incident acoustic waves ($\alpha_m \approx 1$). Approximately equally strong echos came from the chamber floor.

Training histograms were based on 2500 speech frames. Measurement histograms comprised an ensemble of time delays derived from 225 measurement frames. Each frame was $L = 800$ samples long. A male speech signal was used for training, and a female voice was used for measurements. For training, a speaker was placed at the center of the P angular regions. When performing measurements, the speaker was positioned arbitrarily. We counted how many times the speaker positions were correctly assigned to the underlying angular region. This number was used to approximate the probability of correctly estimated speaker orientations, P_D, according to

$$P_D \approx \frac{\text{Number of correct estimates}}{\text{Total number of estimates}}. \qquad (15)$$

Three source localization techniques were compared:

1. No classification, i.e., Eq. (1) was used to estimate $\theta^{(p)}$.
2. ML classification according to Eq. (11), and
3. χ^2-classification as outlined in Eq. (14).

Due to the design of the algorithms, a different number of frames is necessary to arrive at the number of correct angular estimates. When either time delay estimates are used directly to arrive at an angular estimate or when the ML classification technique is applied, we get an instant estimate for each frame. For both cases, we decided to use 100 frames to estimate P_D. On the other hand, using the same number of source positions to evaluate histogram classification, we need $100 \times 225 = 22500$ frames, since each angular estimate requires a measurement histogram accumulated over 225 frames.

Results are displayed in Fig. 4. The chart on the left side shows the percentage of correctly estimated incident angles when there are strong acoustic reverberations but no noise. We see that there is no difference between the traditional technique (associated with the first bar) and ML time delay classification whose result is shown in the second bar. Both obtain the same $P_D = 55\%$. Much better results are obtained by histogram classification. As indicated by the third bar, it achieves $P_D = 92\%$. Estimating the angle of incidence using histogram classification, thus, provides an average improvement in localization accuracy of 67% over the first two methods.

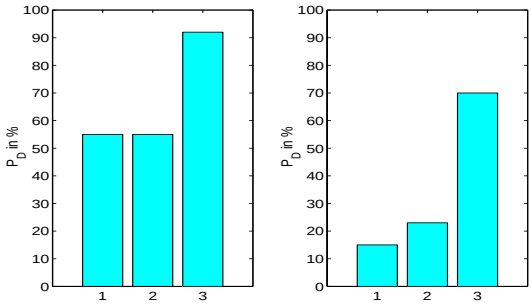

Figure 4: Probability, P_D, of correct source localization when only echos are present (left) and with echos, background noise, and an acoustic obstruction (right). The indices along the horizontal axis refer to the traditional method (1), the ML classification (2), and the classification by histogram comparison (3).

The chart on the right side of Fig. 4 displays results when the direct path from the source to the array is no longer accessible and when there is a background noise source with a signal-to-noise-ratio of 6 dB. As before, strong echos are generated by reverberations from the left and right wall and the floor. This time, the second bar shows that the ML classification method achieves only $P_D = 23\%$. The traditional speaker localization method obtains an even lower $P_D = 15\%$ as indicated by the first bar. Nevertheless, using ML classification, it still remains somewhat possible to distinguish between source directions even when the traditional technique fails. The third bar reveals that histogram classification still yields a superior $P_D = 70\%$.

5. DISCUSSION AND CONCLUSIONS

We presented an approach to robust speaker localization under adverse acoustic conditions. It is based on the classification of time delays carried out using estimates of conditional probability density functions. Two methods were investigated in detail: ML classification and classification by histogram comparison.

Experimental results based on real data in adverse acoustic environments were performed. The results indicate that source localization by ML classification performs approximately as poorly as the traditional method using straightforward geometrical relationships between time delay estimates and speaker positions. Speaker localization based on histogram comparison, however, yields much better results. Given our experimental setup, it outperformed the ML method by 67% when echos were present. In another experiment where we additionally considered an acoustic obstruction and ambient background noise, it delivered results beating the remaining two techniques by an even wider margin.

The excellent results obtained by histogram classification carry the price tag of longer observation periods. For example, creating a measurement histogram using 200 frames, each containing 512 samples acquired with a sampling rate of $f_s = 48\,\text{kHz}$, takes approximately two seconds. This time can, however, be halved, if segments are overlapped by a factor of two. Nevertheless, if (1) a statistical description of the acoustic environment is accessible, and (2) system specifications permit the time to evaluate an ensemble of time delay estimates, then speaker localization based on histogram comparison is among the most robust methods suggested so far.

6. ACKNOWLEDGMENTS

This work is part of the ongoing Sonderforschungsbereich (SFB) No. 603 being carried out at the University of Erlangen-Nürnberg. It is supported by the Deutsche Forschungsgemeinschaft (DFG). The authors acknowledge the assistance of Stefan Kempf and advice by Joachim Hornegger.

7. REFERENCES

[1] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320 – 327, 1976.

[2] Y. T. Chan and K. C. Ho. A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, 42(8):1905–1915, 1994.

[3] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 5(1):45 – 50, 1997.

[4] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 187–90, Munich, 1997.

[5] M. Omologo and P. Svaizer. Acoustic source location in noisy and reverberant environment using CSP analysis. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 901 – 904, 1996.

[6] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(4):943 – 949, 1979.