ICSV14
Cairns • Australia
9-12 July, 2007

LOCALIZATION AND IDENTIFICATION OF PERSONS AND AMBIENT NOISE SOURCES VIA ACOUSTIC SCENE ANALYSIS

Alexej Swerdlow, Kristian Kroschel, Timo Machmer, Dirk Bechler

Institut für Nachrichtentechnik (INT)
Karlsruhe Institute of Technology (KIT), Universität Karlsruhe (TH)
Kaiserstr. 12, 76128 Karlsruhe
Germany

{swerdlow,kroschel,machmer}@int.uni-karlsruhe.de

Abstract

There are diverse areas of application for the acoustic scene analysis, consisting of localization and identification of acoustically observable sound sources. In particular, the man-machine interaction in the broadest sense is of peculiar interest. In this paper a method for the passive acoustic localization of sound sources using time difference of arrival (TDOA) estimates in microphone pairs as well as an approach for the classification of ambient noise sources, based on autoregressive (AR) models, are presented. Therewith, classification of individual sound source categories is possible, although their spectral characteristics can vary significantly.

1. INTRODUCTION

There are a lot of areas, in which acoustic scene analysis is required. One of the most important is the interaction between man and machine, which is given in scenarios, where a human interacts with a machine, for example a so called *humanoid robot*, or is assisted by one. Normally, the communication takes place via speech. In this case it is important for the robot to know, who the speaking person is and where he stands. In situations when no immediate contact between the user and the machine takes place, many other active sound sources can still exist in the robots proximity. A common example is a kitchen, which contains many different appliances that can be acoustically observed in most cases. The robot ought to know its environment at any time to be able to find its way around. Especially, if handicapped or elderly people are involved, the humanoid robot has to guarantee the security of these people. Due to reduced ability to hear, an elderly person might not register that the telephone rings, so that the humanoid robot has to give a hint concerning this event. Thus, the humanoid robot has to compensate the deficiency to hear of the person, which the robot takes care of.

The man-machine interaction within a vehicle is another example for abilities of an acoustic based scene analysis. Thereby, the users and their positions within the vehicle are of peculiar interest. If the specific seat, from where the car is controlled, and the operating person are

known, it will be possible to parameterize the selected control instructions with some position specific properties. Demonstrative examples are the seat and air conditioning settings within the vehicle, or manipulations of infotainment systems.

Thus the acoustic scene analysis consists of two domains: localization of sound sources and their classification, or identification respectively. Some approaches covering both fields of research are presented in the sequel.

2. SOUND SOURCE LOCALIZATION

The technique of choice in most passive acoustic sound source localization systems using a microphone array is a two-step procedure. First, the time difference of arrival (TDOA) of sound signals in a pair of spatially separated microphones is estimated. Then the estimated TDOA in combination with the known microphone array geometry is used for the localization of the sound source in the environment.

2.1. Signal Model

For a given pair of spatially separated microphones M_i and M_j, the microphone signals $x_i(t)$ and $x_j(t)$ for a source signal $s(t)$, propagated through a noisy and reverberant environment, can be modelled mathematically as

$$x_i(t) = h_i(t) * s(t) + n_i(t) \tag{1}$$

$$x_j(t) = h_j(t) * s(t - \tau_{ij}) + n_j(t), \tag{2}$$

where τ_{ij} represents the relative signal delay of interest, $*$ signifies the convolution operator, $h_i(t)$ is the acoustic impulse response between the sound source and the i^{th} microphone, and the additive term $n_i(t)$ summarizes the channel noise in the microphone system as well as environmental noise for the i^{th} sensor. This noise $n_i(t)$ is assumed to be uncorrelated with $s(t)$ and $n_j(t)$. The TDOA estimation attempts to compute τ_{ij} of the direct-path time delays τ_i and τ_j of the microphone signals $x_i(t)$ and $x_j(t)$, defined as

$$\tau_{ij} = \tau_j - \tau_i. \tag{3}$$

2.2. TDOA Estimation with the GCC Method

The most popular approach for determining the TDOAs is the Generalized Cross Correlation (GCC) method, presented by Knapp and Carter [1]. The relative time delay τ_{ij} is estimated as the time lag with the global maximum peak in the GCC function $R_{ij}^{(g)}(\tau)$:

$$\hat{\tau}_{ij} = \arg\max_{\tau} R_{ij}^{(g)}(\tau). \tag{4}$$

This GCC function $R_{ij}^{(g)}(\tau)$ is defined as

$$R_{ij}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau}\, d\omega \tag{5}$$

with $X_i(\omega)$ the Fourier transform of $x_i(t)$.

The weighting function ψ_{ij} intends to decrease noise and reverberation influence and tries to emphasize the GCC peak at the true TDOA τ_{ij}. For real environments, the *Phase Transform (PHAT)* technique has shown the best performance [2]. The PHAT weighting function is defined as

$$\psi_{ij}^{PHAT}(\omega) = \frac{1}{|X_i(\omega)X_j(\omega)^*|}, \tag{6}$$

and can be regarded as a whitening filter.

2.3. Reliability Criterion for TDOA Estimates

Although the GCC approach seems to be practical, its application in real acoustic environments is only of limited use. Even in mildly reverberant rooms, the TDOA estimation error rate rises significantly, delivering unreliable time delays and hence non-confident sound source locations. Therefore, reliability indicators are required allowing to evaluate the confidence of every single TDOA estimate.

As we showed in the past [3], the absolute value of the first maximum peak in the GCC function can be used very efficiently to evaluate the reliability of the actual TDOA estimate. This criterion allows a reliability scoring of individual estimates and can be used to reject erroneous measurements. The higher the value of the first peak in the GCC function is, the higher is the probability that the TDOA was estimated correctly.

3. SOUND SOURCE IDENTIFICATION

In addition to the acoustic localization, the identification of localized persons and ambient noise sources is another major part of the acoustic scene analysis. Besides forensic applications, the interaction between man and machine gains more and more importance. Typical applications are for instant the identification of speakers by humanoid robots or the identification of passengers within a vehicle to adjust position and speaker specific properties.

Therefore two different approaches are presented below. We use the Mel Frequency Cepstral Coefficients (MFCC) as features in combination with the Gaussian Mixture Model (GMM) to identify speakers. For classification of ambient noise sources that occur within earshot, a method, which applies linear prediction based on the autoregressive (AR) models, was developed.

3.1. Text-independent Speaker Identification

The Mel Frequency Cepstral Coefficients (MFCC) have proven to be the most appropriate parameters for speaker identification [4], which are also used as basic features for speech recognition. The sampled instationary speech signal $s(k)$ requires a short time spectral analysis based on segments of 16 ms each, within which the signal is assumed to be stationary. These segments with an overlap of the factor 0.5 and weighted with a Hamming window are transformed into the frequency domain by FFT of length $N = 256$. Using the Mel filter bank [5], which is similar to the spectral selectivity of the human ear, a reduced spectral representation is found by 40 filters with a triangular spectral shape. Below 1 kHz, 13 filters are spaced equally, whereas the other 27 filters are spaced logarithmically along the frequency axis. The logarithm of the output of the 40 filters is applied to the Discrete Cosine Transform (DCT), which decorrelates the parameters. The 13 largest of these parameters form the MFCC vector of the analyzed speech

segment. The corresponding statistical speaker model as well as a real-time demonstrator were presented by Kroschel [6].

3.2. Ambient Noise Source Identification

For the classification of ambient noise sources, we present another approach. Like speaker identification, these sources are usually instationary. That is why the sampled sound source signal $s(k)$ requires a short time spectral analysis based on segments of 16 ms and an overlap of the factor 0.5. Data processing takes place in the time domain, in contrary to the speaker identification.

3.2.1. Event Detection

In order to be able to detect an acoustic event, the energy within a frame is calculated for each frame. The energy $en(\kappa)$ in the frame κ of length $N = 256$ is defined as

$$en(\kappa) = \frac{1}{N} \sum_{k=n_\kappa}^{n_\kappa+N-1} s(k)^2 \tag{7}$$

with n_κ the number of the sample, which is the first one in the frame κ. The weighting with the frame length is done to get a frame length independent rate for the energy. An acoustic event is detected, as soon as $en(\kappa)$ exceeds a previously defined threshold value e_{on} and ends, when $en(\kappa)$ falls below another energy threshold value e_{off}.

3.2.2. Classification with AR models

For the classification of acoustic events, autoregressive (AR) models are used. For each sound class $K^{(c)}$ with $c = 1, ..., N_k$ to be recognized, one or more AR models $\mathbf{p}_j^{(c)}$ with $j = 1, ..., P^{(c)}$ of order M are appointed. For every sound class $K^{(c)}$ and the associated prediction coefficients $\mathbf{p}_j^{(c)}$, the prediction error $e_j^{(c)}(k)$ for the sample $s(k)$ is determined in the following way:

$$e_j^{(c)}(k) = s(k) - \sum_{\ell=1}^{M} p_{j,\ell}^{(c)} s(k-\ell). \tag{8}$$

To be able to determine, which model fits the currently handled frame κ at best, the energy of the prediction error signal $\epsilon_j^{(c)}(\kappa)$ is calculated for every sound class $K^{(c)}$ and the associated models $\mathbf{p}_j^{(c)}$ over the entire frame:

$$\epsilon_j^{(c)}(\kappa) = \sum_{i=n_\kappa}^{n_\kappa+I-1} e_j^{(c)}(k)^2. \tag{9}$$

Subsequently, the value of the prediction error of the model $\mathbf{p}_j^{(c)}$ and the sound class $K^{(c)}$, which represents the frame κ at best, is then defined by

$$\epsilon_{min}^{(c)}(\kappa) = \min_{j=1,...,P^{(c)}} \epsilon_j^{(c)}(\kappa). \tag{10}$$

Finally, the frame κ is assigned to the estimated noise source class $\hat{K}$ in the following way:

$$\hat{K}(\kappa) = \underset{c=1,\ldots,N_K}{\arg\min}\ \epsilon_{min}^{(c)}(\kappa). \tag{11}$$

In order to classify the current acoustic event, frames are aggregated into blocks of defined size. A trade-off has to be made between a high percentage of correct classification results and a high number of estimates, which is crucial for the continuous real-time classification. The entire acoustic event within the actual block is matched to the noise source class, which prevails in this block.

4. EXPERIMENTAL SETUPS AND SELECTED RESULTS

For data recording, omni-directional electret condenser microphones were used. Real experiments were carried out in different test environments. Investigations were examined in a typical office room as well in an exemplary up to date car. The distance of the microphone pairs for localization with the GCC method were varied between 20 cm (concentrated microphone array in the head of a humanoid robot) and 1.14 m (distributed microphone array in a car).

4.1. Evaluation of the Reliability Criterion for TDOA Estimates

To determine the relationship between the the maximum peak of the GCC function and the TDOA reliability, TDOA estimates were divided into 15 intervals. The interval borders are extracted from the histogram for the maximum peak of all analysis frames (Figure 1). The interval limits were chosen such that every interval contains a similar number of TDOA estimates. Different utterances of German sentences (altogether 47850 words) from 6 speakers (3 male and 3 female) were played back by a loudspeaker, which was placed in 25 different positions in an office room of 5m x 5m x 3m with typical environmental noise (SNR $\approx$ 19 dB) coming from fans, mechanical equipment, etc. and relatively strong reverberations (reverberation time $T_{60} \approx$ 350 ms). Table 1 details the interval borders. It also shows the correct estimate percentage per

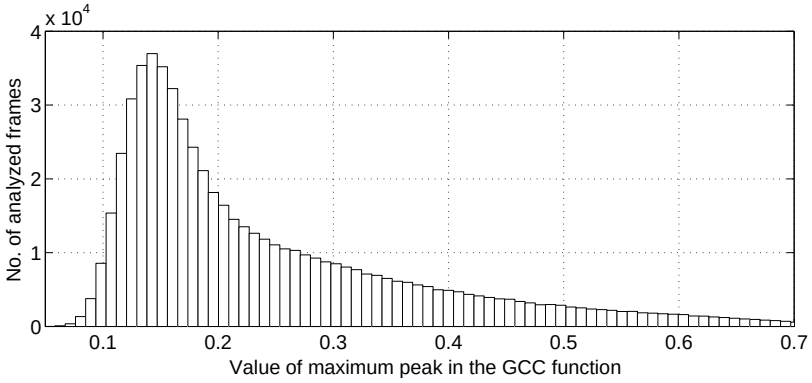

Figure 1. Histogram for the maximum peak criterion values of all analysis frames.

interval for increasing values of the maximum peak, exemplarily for a concentrated array of 5 microphones in an equilateral double-tetrahedron geometry with a side length of 28 cm, A TDOA estimation is deemed correct, if the product of the sampling frequency f_s and the term

$|\hat{\tau}_{ij} - \tau ij|$, i.e. the absolute value of the difference of the estimated and the real TDOA value of the sound source, is less than a decision threshold of $T_{dec} = 1.5$ samples

$$f_s \cdot |\hat{\tau}_{ij} - \tau_{ij}| \begin{cases} \leq T_{dec} & : \text{ correct} \\ > T_{dec} & : \text{ false.} \end{cases} \tag{12}$$

As can be seen, the maximum peak in the GCC function allows very convincingly a judgment about the reliability of the current TDOA estimate. Low criterion values mean low reliability of only 15.62% for the maximum peak criterion in interval 1, whereas for high values of the criterion the confidence increases to almost 100%, delivering highly reliable estimates. Consequently this property of the GCC function can be used to detect outliers and to suppress real environment influences such as noise and room reverberation considerably. With the confidence criterion, a trade-off has to be made between a high number of estimates, which is necessary for a continuous target tracking, and a high percentage of correct TDOA estimates, which is crucial for the robust source localization.

Table 1. Interval borders of the reliability criterion values maximum peak (m) and correct estimate percentage per interval.

Interval	**Maximum peak m**	**Correct estimate percentage**		**Interval**	**Maximum peak m**	**Correct estimate percentage**
1	$m \leq 0.100$	15.62%		9	$0.250 \leq m \leq 0.275$	94.76%
2	$0.100 \leq m \leq 0.120$	18.59%		10	$0.275 \leq m \leq 0.300$	96.87%
3	$0.120 \leq m \leq 0.140$	24.26%		11	$0.300 \leq m \leq 0.350$	98.17%
4	$0.140 \leq m \leq 0.160$	32.55%		12	$0.350 \leq m \leq 0.400$	99.20%
5	$0.160 \leq m \leq 0.180$	45.37%		13	$0.400 \leq m \leq 0.500$	99.71%
6	$0.180 \leq m \leq 0.200$	62.54%		14	$0.500 \leq m \leq 0.600$	99.86%
7	$0.200 \leq m \leq 0.225$	79.18%		15	$m \geq 0.600$	99.88%
8	$0.225 \leq m \leq 0.250$	90.25%				

4.2. Evaluation of the Ambient Noise Source Identification System

Various kitchen appliances[1] in combination with two untrained sound sources[2] were used for the real-time classification of ambient noise sources. The percentage of correct frame classifications and the needed number of AR models of order 16 for each ambient sound source state are summarized in Table 2. The standard deviation is given in Table 3.

As can be seen, the classification with AR models is a multiple detection issue. That is the reason why also untrained sound sources (speech, knocking noise) are always classified. To avoid this deficiency, a reject class was defined, additionally to the block aggregation described in 3.2.2. A block is rejected in case less than 60 percent of frames within the block classify the

[1] KC(P): kitchen clock (programming), KC(E): kitchen clock (expiration), CG(A): coffee grinder (activity), T(D): toaster (down), T(U): toaster (up), T(U): telephone (ringing), EWJ(H): electric water jug (heating), EWJ(B): electric water jug (boiling)

[2] US(S): untrained source (speech), US(KN): untrained source (knocking noise)

Table 2. Percentage results of the frame based classification with AR models of order 16 for kitchen appliances.

Sound class\AR model	KC(P)	KC(E)	CG(A)	T(D)	T(U)	T(R)	EWJ(H)	EWJ(B)
KC(P)	**98.89**	1.11	0	0	0	0	0	0
KC(E)	1.31	**98.69**	0	0	0	0	0	0
CG(A)	1.03	0	**57.98**	12.59	3.56	0.12	12.99	11.72
T(D)	0.99	0	0.99	**83.52**	11.52	0	2.34	0.63
T(U)	0.99	0	0.59	12.04	**86.1**	0	0.28	0
T(R)	0.99	0	0.51	0.24	0.16	**92.32**	3.60	2.18
EWJ(H)	0.99	0	1.47	5.27	0.51	0.20	**87.60**	3.96
EWJ(B)	1.78	0	0.40	0.36	0.12	0.08	2.06	**95.21**
US(S)	1.70	0	21.43	2.53	2.61	2.02	32.75	36.95
US(KN)	0.99	0.04	9.47	31.49	12.83	0.44	43.92	0.83
Average number of needed AR models	**5.60**	**5.40**	**16.44**	**17.60**	**16.56**	**14.88**	**17.04**	**21.28**

Table 3. Standard deviation for the frame based classification matrix with AR models of order 16 for kitchen appliances.

Sound class\AR model	KC(P)	KC(E)	CG(A)	T(D)	T(U)	T(R)	EWJ(H)	EWJ(B)
KC(P)	**0.82**	0.82	0	0	0	0	0	0
KC(E)	0.30	**0.30**	0	0	0	0	0	0
CG(A)	0.09	0	**3.95**	3.61	1.36	0.18	1.88	4.20
T(D)	0	0	0.46	**3.05**	2.89	0	0.57	0.33
T(U)	0	0	0.82	3.42	**3.12**	0	0.18	0
T(R)	0	0	0.65	0.43	0.22	**2.50**	1.94	0.66
EWJ(H)	0	0	0.82	2.13	0.64	0.34	**1.01**	1.51
EWJ(B)	0.90	0	0.40	0.38	0.18	0.11	0.74	**1.07**
US(S)	0.54	0	3.30	0.53	0.98	1.59	2.67	3.29
US(KN)	0	0.09	2.69	2.63	2.01	0.35	3.29	0.33

same sound class. One block consists of 62 frames, so that acoustic segments with the length of approximately one second were analyzed. Percentage results for the block based classification are presented in Table 4.

It is visible, that using the presented approach, which is based on autoregressive (AR) models, the classification of individual sound source categories is feasible, although their spectral characteristics vary significantly. Noise sound classes, which differentiate in their reproducibility, are difficult to classify. This is true for instance for the coffee grinder. An improvement could be achieved by increasing the number of AR models, but this would also raise the calculating costs significantly.

ACKNOWLEDGMENT

This work has been supported by the German Science Foundation DFG within the Sonderforschungsbereich 588 *Humanoid Robots*.

Table 4. Percentage results of block based classification with AR models of order 16 for kitchen appliances and a reject class for untrained noise sources.

Sound class\AR model	**KC(P)**	**KC(E)**	**CG(A)**	**T(D)**	**T(U)**	**T(R)**	**EWJ(H)**	**EWJ(B)**	**Reject**
KC(P)	**96.67**	0	0	0	0	0	0	0	3.33
KC(E)	0	**100.00**	0	0	0	0	0	0	0
CG(A)	0	0	**69.33**	1.33	2.67	0	5.33	0	21.33
T(D)	0	0	0	**78.67**	1.33	0	0	0	20.00
T(U)	0	0	0	0	**98.67**	0	0	0	1.33
T(R)	0	0	0	0	0	**98.67**	0	0	1.33
EWJ(H)	0	0	0	0	0	0	**89.33**	4.00	6.67
EWJ(B)	0	0	0	0	0	0	2.00	**96.00**	2.00
US(S)	0	0	4.00	0	1.33	0	12.67	8.00	**74.00**
US(KN)	0	0	0.67	0	0.67	0	41.33	0	**57.33**

REFERENCES

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

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. *Robust localization in reverberant rooms*, chapter 8, pages 157–180. Springer, Berlin, 2001.

[3] D. Bechler and K. Kroschel. Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays. In *13. Konferenz Elektronische Sprachsignalverarbeitung ESSV*, Dresden, 2002.

[4] D. O'Shaughnessy. *Speech Communication - Human and Machine*. IEEE Press, New York, 2000.

[5] B. Gold and N. Morgan. *Speech and Audio Signal Processing*. Wiley, New York, 2000.

[6] K. Kroschel and D. Bechler. Demonstrator for automatic text-independent speaker identification. In *DAGA 2006*, Braunschweig, 2006.