Position Estimation of Car Occupants by Means of Voice Analysis

T. Machmer[1], A. Swerdlow[1], B. Kühn[1], K. Kroschel[2]
[1] *University of Karlsruhe, Institute for Anthropomatics, 76128 Karlsruhe, Germany,*
Email: {machmer, swerdlow, kuehn}@ies.uni-karlsruhe.de
[2] *Fraunhofer-Institute for Information and Data Processing, 76131 Karlsruhe, Germany,*
Email: kristian.kroschel@iitb.fraunhofer.de

Introduction

In the next-generation cars, the interaction between man and machine via speech recognition gains more and more in importance. Especially for the convenient control of comfort and entertainment functions, this kind of interaction is already implemented in current upper class vehicles. A specific property of all these systems entails having to activate the system before using, i.e. the user has to press a button (mostly positioned on the steering wheel) before he can start giving voice commands. By pressing such an activation button, two main tasks are fulfilled. On one side, there is no need for a separated Voice Activity Detection (VAD) system. On the other side, it is ensured that the command originates from a seat position from which the activation button can be reached.

The goal of our work is the optimization of the second task. Therefore, a microphone array which is mounted in the car, is used. We endeavour to gain the information about the position of car passengers by evaluating their common speech activity and in so doing to replace the necessity of the explicit activation of the system by pressing a button. Besides this, the functionality of the system can be extended to all passengers inside the car and must not be restricted to the driver seat only.

In recent years, our research group developed a real-time system called CarOPE (**Car O**ccupants **P**osition **E**stimation) for position estimation of speaking car passengers. During the years, we optimized the array geometry and the algorithms. The achieved results were presented on DAGA conferences in 2007 and 2008 [1, 2].

The current paper evaluates an improved correlation based localization method and compares it to an intensity based approach. Furthermore, two new close-to-production array geometries are evaluated: a linear array placed near the rear-view mirror as well as a distributed array with microphones placed above four passenger seats. All algorithms and geometries are evaluated with real world data under realistic driving conditions. Additional to different driving situations, car-internal noise sources like the fans of the air conditioning system, the audio system, or an opened window are investigated.

Correlation Based Localization

The first localization approach we used is based on the estimation of the time difference of arrival (TDOA) of sound signals in a pair of spatially separated microphones. The most common technique for the determination of TDOAs is the generalized cross correlation (GCC) [3]. The GCC function $R_{ij}^{(g)}(\tau)$ is defined as

$$R_{ij}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}^{PHAT}(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau}\, d\omega, \quad (1)$$

where $X_i(\omega)$ is the Fourier-Transform of the given microphone signal x_i. ψ_{ij} is a weighting function which intends to decrease environmental influences and tries to emphasize the GCC peak at the true TDOA. For real environments, the Phase Transform (PHAT) technique has shown the best performance [3]. The PHAT weighting function is defined as

$$\psi_{ij}^{PHAT}(\omega) = \frac{1}{|X_i(\omega) X_j(\omega)^*|} \quad (2)$$

and can be regarded as a whitening filter.

The relative time delay between the microphones τ_{ij} is estimated as the time lag with the global maximum peak in the GCC function $R_{ij}^{(g)}(\tau)$:

$$\hat{\tau}_{ij} = \arg\max_{\tau} R_{ij}^{(g)}(\tau). \quad (3)$$

As shown in [4], 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.

The estimation of the seated position is done by a hierarchical analysis of the estimated time delays $\hat{\tau}_{ij}$ and the maximum correlation peak value in the microphone pairs of the used array.

Intensity Based Localization

In order to estimate the seat position of the speaking person, the noise energy floor $P_i^{(n)}$ is estimated for each microphone i by means of a minimum statistics approach, inspired by [5]. Using this noise estimation, the signal-to-noise ratio (SNR) is calculated by

$$SNR_i = 10 \cdot log\left(\frac{P_i}{P_i^{(n)}}\right), \quad (4)$$

where P_i is the signal energy in the microphone i.

In order to determine the current active seat, the seat position with the maximum SNR is searched.

Similar to the the correlation based localization approach, a reliability criterion for the intensity based method was found. Therefore, the estimation of the SNR value of the winning seat is used to reject unreliable position estimations. In doing so, a higher SNR value indicates a probably more reliable estimation.

Experimental Setup

All experimental recordings were carried out in an exemplary up-to-date car using two different microphone arrays, which were mounted inside the vehicle. The ceiling array consists of four distributed microphones C_1-C_4 above the passenger seats. The second array is a linear array placed near the rear-view mirror and consists of four microphones L_1-L_4 with an inter-microphone distance of 2.5 cm. Figure 1 shows the microphone positions inside the car in a schematic overview.

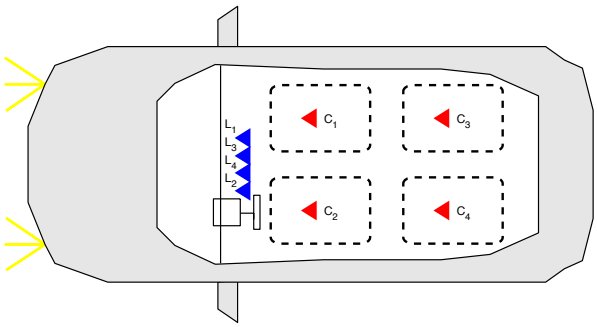

Figure 1: Schematic illustration of the experimental setup with linear array microphones L_1-L_4 and ceiling microphones C_1-C_4.

In total, three different localization methods were used. As mentioned above, there are two alternative localization approaches: correlation and intensity based. Both were combined with the ceiling array (C4C: correlation, C4I: intensity), whereas the linear array was used in combination with the correlation based method (L4C) only.

We collected 5.5 hours speech data in total from four speakers (one female, three male). Thereby, the data set was divided into two main parts. The first part consists of recordings done in a parking car with the running engine. Additionally and in order to evaluate car internal influences, we took recordings with the air conditioning system, which was set to the maximal level, as well as recordings with music in the background. The source for the music was an audio CD, whereby the volume was set to the level, which made a conversation between two persons possible. The second part of the speech database consisted of recordings, which were done in typical real driving situations, e.g in a city, on a country road, or on a freeway.

The recorded speech was analyzed in frames of approximately 170 ms. For the data segmentation, a Hamming window with a 50% overlap was applied. We used a sampling frequency of 24 kHz for the ceiling array, in contrast to 96 kHz for a higher temporal resolution in combination with the linear array.

Evaluations in Real Scenarios

Car Internal Influences

In order to guarantee a fair comparison of all algorithms and array geometries, we used recordings in the parking situation and applied a simple frame energy based Voice Activity Detection (VAD) approach. After the application of this VAD to the data, only frames with speech remained.

The localization results for all three methods are shown in Table 1. An estimation of a speaker position is deemed correct if the calculated position is located within the proper seat of the speaker. The localization estimations under the method reliability threshold are rejected. All results are frame based and given in percent, i.e. a correct localization rate of 100 % means that the seat position estimation was correct for all available frames.

It can be seen that the localization rates depend on both the array geometry and the used algorithm. Comparing the two localization methods for the ceiling array (C4C, C4I), it can be seen that both the correlation and the intensity based localization method perform nearly equal for all seat positions, in contrast to the method L4C which performs well on the two front seats only and shows a significant loss of the localization rate for the rear seats.

localization method		**driver**	**co-driver**	**rear left**	**rear right**
C4C	*correct*	91.48	84.91	96.47	95.31
	false	0.49	0.67	0.52	0.17
	reject	8.03	14.42	3.01	4.52
L4C	*correct*	93.08	82.91	45.88	75.58
	false	0.12	0.03	0.16	0.42
	reject	6.79	17.06	53.96	24.00
C4I	*correct*	61.95	63.58	66.19	69.83
	false	0.90	0.08	4.18	0.25
	reject	37.15	36.34	29.62	29.92

Table 1: Localization rates in percent (averaged over four test persons, two minutes of test data per person and seat position) for three localization methods in the parking scenario with VAD.

In order to evaluate the influence of car internal noise sources (air conditioning system, audio system), additional recordings were completed. It has to be mentioned that these evaluations were also performed with VAD. Unlike the above results, the influence measurements were carried out with only one person. The corresponding averaged results for all four seat positions are summarized in Table 2.

It can be seen that the localization rate decreases when one of the noise sources is active. However, only a slight influence can be noticed.

localization method		no noise	audio system	fan max
C4C	*correct*	96.47	91.92	85.31
	false	0.18	0.33	1.59
	reject	3.35	7.75	13.10
L4C	*correct*	92.08	69.86	78.59
	false	0.07	2.02	2.06
	reject	7.85	28.12	19.35
C4I	*correct*	78.70	69.10	52.07
	false	0.42	0.10	0.71
	reject	20.88	30.80	47.22

Table 2: Localization rates in percent (averaged over four seat positions of one test person, four minutes of test data per situation) for three localization methods in the parking scenario with VAD and different car internal influences.

External Noise Sources

For all the following evaluations, we did not use a VAD. Instead of the VAD, the reliability thresholds of the localization methods were used in order to reject unreliable localization estimations. This way of proceeding reflects the activity detection of the CarOPE system and makes measurement results comparable to the performance of the real-time system. In doing so, it has to be mentioned that the rejected localizations contain frames with no speech activity (typically about 10 %) as well as estimations, which were rejected due to the reliability threshold.

At first, the influence of different driving situations to the system accuracy was investigated. Table 3 shows the corresponding results. Through the situations, the driving speed as well as the noise level increases permanently. As it can be expected, the correct localization rates decrease constantly from the parking scenario to the freeway scenario. However, the false localization rate remain low in all scenarios, whereas reject localization rate increases.

localization method		parking	city	country road	free-way
C4C	*correct*	86.12	76.28	70.71	58.01
	false	0.99	2.16	1.98	1.41
	reject	12.90	21.56	27.31	40.58
L4C	*correct*	68.35	54.88	61.87	56.98
	false	0.17	1.74	1.84	2.30
	reject	31.48	43.38	36.29	40.72
C4I	*correct*	81.77	66.44	45.18	28.81
	false	1.52	4.19	4.60	2.90
	reject	16.71	29.37	50.21	68.29

Table 3: Localization rates in percent (averaged over four test persons and four seat positions, two minutes of test data per person for each seat position) for three localization methods in different driving situations without VAD.

In addition to different driving situations, the influence of the internal noise sources was also evaluated in the freeway scenario and without VAD. Besides the the air conditioning system and the audio system, the influence of an opened window (about 10 cm) on the localization accuracy was investigated. Table 4 summarizes the corresponding results.

It can be seen, that the audio system as well as the air conditioning system have similar influence on the localization accuracy comparable to the results in the parking scenario given in Table 2. However, the false localization increases dramatically for the localization method L4C and slightly for C4I.

Nevertheless, the correlation based method in combination with the ceiling array (C4C) achieves reliable seat position estimations for all scenarios.

localization method		no noise	audio system	fan max	opened window
C4C	*correct*	75.98	73.34	57.12	63.52
	false	0.39	5.08	0.17	0.34
	reject	23.64	21.59	42.70	36.13
L4C	*correct*	78.05	57.08	39.08	59.04
	false	0.89	5.20	3.97	8.66
	reject	21.06	37.71	56.95	32.30
C4I	*correct*	53.84	57.59	47.91	57.25
	false	1.16	1.28	4.31	20.95
	reject	45.00	41.13	47.78	21.80

Table 4: Localization rates in percent (averaged over four seat positions of one test person, four minutes of test data per situation) for three localization methods in the freeway situation without VAD and in combination with different car internal influences.

Conclusion

Speaker position estimation in vehicles by means of the voice analysis is a promising task. Especially the correlation based approaches showed very reliable results in all situations. In this paper, we presented an in-car localization system called CarOPE. It makes possible to determine the seated positions of car occupants. Due to its real-time capability, the system could be already successfully integrated in a real car environment.

Acknowledgment

This work has been supported by the German Science Foundation DFG within the The Collaborative Research Center 588: *Humanoid Robots - Learning and Cooperating Multimodal Robots.*

References

[1] Swerdlow A., Kroschel K., Machmer T.: Speaker localization in vehicles via acoustic analysis. DAGA 2007, Stuttgart, Germany, 2007

[2] Swerdlow A., Machmer T., Kühn B., Kroschel K.: Speaker position estimation in vehicles by means of acoustic analysis. DAGA 2008, Dresden, Germany, 2008

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

[4] Bechler, D., Kroschel, K.: Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays, 13. ESSV, Dresden, 2002

[5] Martin, R.: Spectral subtraction based on minimum statistics. EUSIPCO, Edinburgh, Scotland, pp. 1182-1185, 1994