A Robust Acoustic Head Orientation Estimation and Speech Enhancement for In-Car Communication Systems

Rasool Al-Mafrachi, Marco Gimm, Gerhard Schmidt

Christian-Albrechts-Universität zu Kiel, 24118 Kiel, E-Mail:[ral,mgi,gus]@tf.uni-kiel.de

Abstract

Due to the acoustic directionality of human head, the orientation of speakers head in a car could affect the level and SNR of speech signals recorded by ICC system microphones, and consequently influences the performance of the technologies deployed based on those signals. Therefore, estimating the head orientation of speaker in car would be useful for applying the corresponding signal processing techniques which required to reduce the degraded intelligibility and distortion in those speech signals. In this contribution we propose two different acoustic head orientation estimation approaches on the basis of multi-microphones recordings and we evaluate their performance in real car environments within a car driven at different speeds on various roads. In addition, we use the estimated head orientation cues towards the enhancement of ICC system speech signals via appropriate equalization filters and a corresponding noise reduction scheme. The performance is evaluated under various conditions and robust promising results are obtained.

1. Introduction

In automotive environments, the communication among car passengers could be very difficult as usually there are high amount of background noise levels when driving at high or even moderate velocities. Therefore, the communication partners will start to increase their vocal effort (the so-called *Lombard effect*) to compensate for this weak signal to noise ratio situation and further may start to change their head orientation and lean towards each other to reduce the distance between them and hence increasing the conversation intelligibility. However, these actions are uncomfortable for long-time conversations and at the same time have safety risks if the driver does such actions. An ICC system improves the signal-to-noise ratio (SNR) within the car compartment by recording, processing, and playing back the desired speech signal of the talking passenger over loudspeakers located close to the listening passengers. Such ICC system usually performs a set of various real-time signal processing techniques in order to overcome several challenges resulting from the closed electro-acoustic loop operation and the high amount of disturbing noises (wind, engine, tire noise, etc.).

The directionality of human head acoustic field creates another challenge for such ICC system, since frequency selective attenuated speech signals are recorded by the ICC microphones when the talking passengers turn their head. According to Brian B. Monsona and Eric J. Hunter [1], the speech signal is attenuated by $5-10$ dB with respect to the head direction of the speaker. For this reason, the passengers dedicated microphones are recording a degraded low quality and less intelligibility speech signals when the talking passengers turn their head and this degradation is directly proportional with the angle of the head orientation (assuming 0° is pointing to the best microphone). Therefore, by estimating the head orientation of the speaking passenger, it would be possible to apply the corresponding equalization and noise reduction techniques to the microphone signals proportionally with the estimated angle to compensate the frequency selective attenuation resulting from the head turning.

Several research have been done to estimate the source orientation for applications such as smart room voice commanded devices [2], robots [3], and speech acquisition in reverberant environments [4]. However, no attentions have been given to estimate the head orientation of the speaker in automotive environment where there are low SNR scenarios and the speech signal buried in mixture of stationary and non-stationary background noises.

In our previous contribution [5], we have presented the first investigations towards a robust acoustic head orientation estimation of speaking passengers for automotive environments within a simulated noisy car compartment. However, we aim in this contribution to implement and evaluate more robust acoustic head orientation estimation approaches for ICC system. Also, we aim on utilizing these head orientation information of the speaking passenger to design and implement a suitable equalization and noise reduction schemes to compensate the attenuation error resulted from the speaker head turning.

The paper is organized as follows: Sec. 2 shows our proposed model approaches in detail by describing the methods have been used to estimate the head orientation and the necessary post processing for equalization and noise reduction of the desired ICC microphone signals. Sec. 3 presents the details of the experimental measurements while the results, discussions and evaluations can be found in Sec. 4. Finally, a short conclusion is shown in Sec. 5.

2. Model Approach

Fig. 1 depicts the complete process of the proposed algorithm.

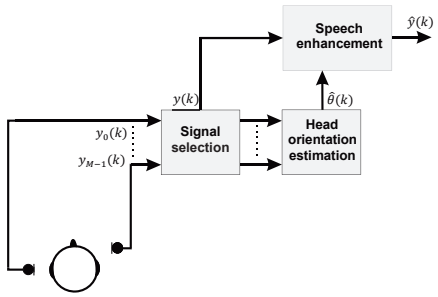


Figure 1: The proposed system top view, k is the time-frame index.

2.1 Signal selection

As it can be seen from Fig.1, the speech of the talking passenger is recorded by a set of microphones ($y_0 - y_{M-1}$) distributed inside the car compartment. Therefore, a signal selection module is necessary to select the required set of microphones signals to be used for the head orientation estimation and at the same time to select the desired microphone signals needed to be further enhanced based on the estimated head orientation angle.

2.2 Head orientation estimation

Fig.2 shows the top view for the head orientation estimation algorithm.

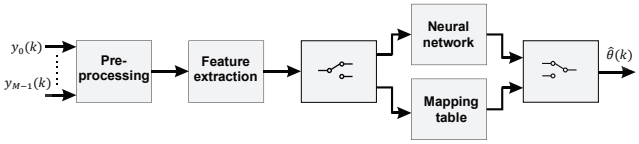


Figure 2: The head orientation estimation algorithm top view.

2.2.1 Pre-processing

The pre-processing is performed in the time-frequency domain $Y_m(k,\mu)$ which is generated by applying an analysis filterbank on the discrete input signal $y_m(k)$, where m, k and μ represent the microphone channel index, frame index and the frequency subband index respectively. For ease of notation, the explanation will be for a single microphone indicated by dropping the subscript m. The parameters for the filterbank in this study are: frame length $= 256$, overlap $= 50\%$, window $=$ Hanning, and the fast Fourier transform length $N_{\mathrm{FFT}} = 256$. A simple threshold based voice activity detector was used to detect the speech presence. The estimated noise $\widehat{N}(k,\mu)$ is compared with the smoothed input magnitude spectrum $\overline{|Y(k,\mu)|}$ multiplied by a certain SNR threshold THR_{SNR}. Whenever the smoothed input magnitude spectrum is larger than the estimated noise reference, the decision flag $\psi(k,\mu)$ is set to 1 otherwise it set to 0. This process is performed framewise for every subband where each flag $\psi(k,\mu)$ represents the decision output at frame k in subband μ.

$$\psi(k,\mu) = \begin{cases} 1, & \text{if } \overline{|Y(k,\mu)|} \cdot THR_{\mathrm{SNR}} > \widehat{N}(k,\mu) \\ 0, & \text{else.} \end{cases} \qquad (1)$$

The SNR threshold THR_{SNR} has been set to -12 dB. The final decision of the VAD for the frame k is determined by comparing the number of active flags $N_{\psi-\mathrm{act}}(k)$ within the frame k against another threshold VAD_{THR} as following:

$$VAD_{\mathrm{decision}}(k) = \begin{cases} 1, & \text{if } N_{\psi-\mathrm{act}}(k) > VAD_{\mathrm{THR}} \\ 0, & \text{else,} \end{cases} \qquad (2)$$

where 1 and 0 represent the presence and absence of the speech respectively. Only the speech frames are preserved which indicated by $\sim$ superscript above the frame index k. Smoothing is applied to reduce the variance within the magnitude of the input speech $|Y(\tilde{k},\mu)|$ and it is performed along the time axis for every subband using a first order IIR filter as following:

$$\overline{|Y(\tilde{k},\mu)|} = \alpha|Y(\tilde{k},\mu)| + (1-\alpha)\overline{|Y(\tilde{k}-1,\mu)|}, \qquad (3)$$

where α is the smoothing constant and equal to 0.8. Fig.3 shows the block diagram of the pre-processing module.

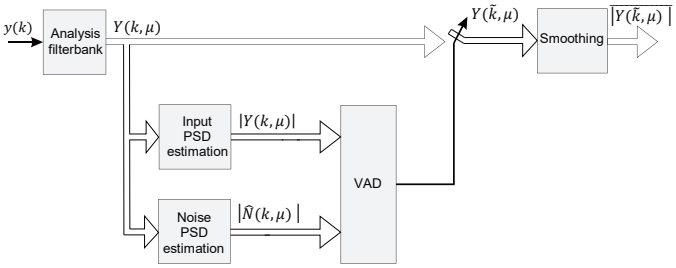


Figure 3: The pre-processing block diagram.

2.2.2 Feature extraction

The power ratio between the microphones signals are used as a feature vector to estimate the head orientation of a speaker. First, the smoothed magnitude of the input speech spectrum $\overline{|Y(\tilde{k},\mu)|}$ for each microphone channel m is filtered with a Mel filterbank in order to reduce the feature vector dimensionality. Fig. 4 shows the block diagram of the feature extraction module.

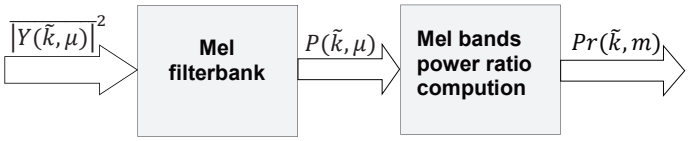


Figure 4: Feature extraction block diagram.

The power ratio are computed for each microphone m and Mel band l as following:

$$Pr(\tilde{k},l) = 10\log_{10}\left[\frac{P(\tilde{k},l)}{\overline{P_M(\tilde{k},l)}}\right], \qquad (4)$$

where $\overline{P_M(\tilde{k},l)}$ is the mean power over the total number of microphone channels M in a given Mel band l which is calculated by:

$$\overline{P_M(\tilde{k},l)} = \frac{1}{M}\sum_{m=0}^{M-1} P_m(\tilde{k},l)\,. \qquad (5)$$

$Pr(\tilde{k},l)$ represents the feature vector which used as the input of the head orientation estimator module. By using the power ratio related feature, we expect that our head orientation estimation approach can model the radiation pattern of the human head and estimate its orientation accurately.

2.2.3 Head orientation estimator

The head orientation estimator module is responsible for mapping the power ratio feature vector into a corresponding head orientation angle. In this study, two versions of the estimator have been implemented. The first version is based on using an artificial neural network (ANN) to estimate the head orientation. The ANN is fully connected in a feedforward configuration with two hidden layers (6 and 3 neurons respectively), both equipped with non-linear sigmoid activation function, and one output layer which equipped with a linear activation function and consist of five neuron corresponding to five angle classes of the head orientation ($-90^\circ, -45^\circ, 0^\circ, 45^\circ, 90^\circ$). The ANN structure should kept simple to reduce the complexity as the ANN after the training

would be used for real time processing, however this is the simplest structure which has been found empirically. The feature vector is divided into three data sets: training 70%, validation 15%, testing 15% and the network is trained using Scaled conjugate gradient backpropagation algorithm.

The second version of head orientation estimator is done using a simple mapping table. In this method fewer number of Mel bands ($L = 4$) have used during the feature extraction in order to reduce the mapping table creation complexity. The power ratio vector is examined manually and a simple mapping table is created by defining a set of power ratio thresholds for each class of head orientation angle. This mapping table is then used to estimate the head orientation angle class. However in this method, only the frames that have a sufficient SNR are used and only three classes of head orientation angles are estimated ($-90^{\circ}, 0^{\circ}, 90^{\circ}$).

2.3 Speech signal enhancement

2.3.1 Equalizer filter design

Two microphones signals are used to design the equalizer filter. The first microphone is mounted near the mouth of the speaker and its spectrum Y_{ref} is used as a reference, while Y is the spectrum of the second microphone (the desired ICC speaker dedicated microphone) which needs to be enhanced. Both microphones signals are recorded for every possible head orientation angles and the corresponding dedicated FIR equalizer filters are designed. The equalization filter design proceeds as follows (see the block diagram in Fig.5):

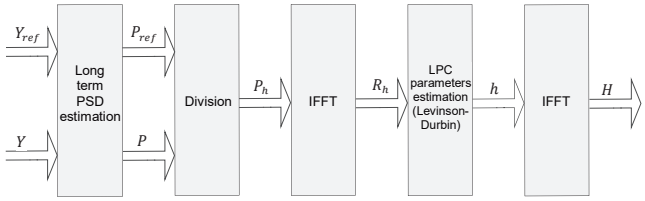


Figure 5: Equalizer filter design process.

Step 1: Estimate the long-term averaged PSDs for both Y_{ref} and Y.

Step 2: Estimate the PSD, P_h of the equalizer filter h.

Step 3: Calculate the autocorrelation sequence R_h of h from P_h.

Step 4: Given the autocorrelation sequence R_h, the coefficients of a p^{th}-order AR (auto-regressive) linear filter h can be estimated by the Levinson-Durbin recursion. This is indeed the equalizer filter we need to correct the selective frequency attenuation in the desired ICC microphone signal which results from head turning with respect to a specific angle θ.

Step 5: The designed filter h is converted to frequency domain H as all the processing in our algorithm is in the time-frequency domain.

2.3.2 Equalization and noise reduction

In this module, a desired ICC microphone signal is equalized based on the estimated head orientation information as it is illustrated in Fig.6.

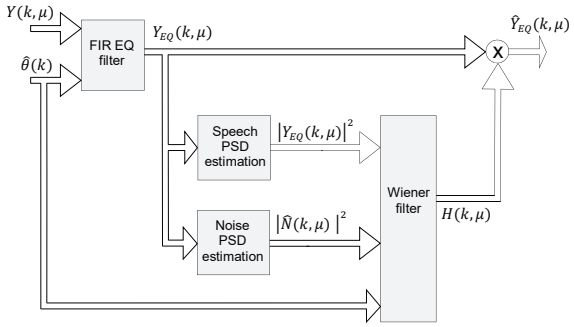


Figure 6: Equalization and noise reduction.

A suitable noise estimation is needed in order to efficiently suppress the background noise which can be increased after applying the equalization filter. The amount of noise floor reduction is directly connected to the estimated angle of the speaker head orientation (the larger the angle the more noise reduction is applied and vice versa). A robust extended noise estimation (RENE), proposed and evaluated in [6], was used. The noise estimation $\hat{N}(k,\mu)$ is modeled as the weighted sum of the smoothed input magnitude spectrum $\overline{|Y_{EQ}(k,\mu)|}$ and the slow changing noise pre-estimator $\hat{N}_{pre}(k,\mu)$ as following:

$$\hat{N}(k,\mu) = \big(1 - w(k,\mu)\big) \cdot \hat{N}_{pre}(k,\mu) + w(k,\mu) \cdot \overline{|Y_{EQ}(k,\mu)|}, \quad (6)$$

where $w(k,\mu)$ is the probability weighting for a speech pause. The estimated noise $\hat{N}(k,\mu)$ is used to compute Wiener filter coefficients $H(k,\mu)$ and an enhanced version of the equalized speech signal spectrum $\hat{Y}_{EQ}(k,\mu)$ is produced as following:

$$\hat{Y}_{EQ}(k,\mu) = Y_{EQ}(k,\mu) \cdot H(k,\mu) \quad (7)$$

3. Experimental measurements

The experimental recordings were conducted inside a car driven at various speeds (0 kmh, 50 kmh, 80 kmh, and 120 kmh) and on various roads (city, regional, and highway). Four native German speakers (3 male and 1 female) were conducted the recording by sitting inside the car and spoke while rotating their head in the azimuth plane towards pre-defined labels inside the car cabin that represent approximately the following angles (from left to right, 0° is the front direction): -90°, -45°, 0°, 45°, and 90°. The speech signals were recorded using four microphones distributed inside the car cabin as illustrated by solid dots at the car interior schematic diagram in Fig.7.

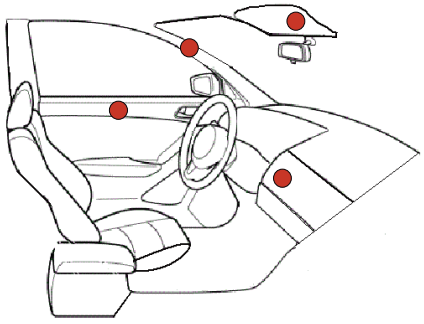
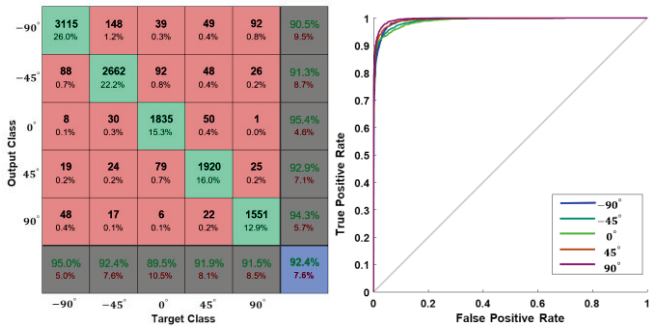


Figure 7: Microphone distribution inside the car cabinet (solid red dots).

4. Results

At first we evaluated the performance of the proposed head orientation estimation models. Fig.8-a shows the confusion matrix that describes the performance of our ANN head orientation estimator model. As it can be seen, the estimator model has a 92.4% accuracy and more than 90% sensitivity and precision for all angle classes which implies that the model is reliable and can estimate the angle of the speakers head orientation accurately under various SNR scenarios. This high accuracy performance can also be evaluated by examining the ROC curves of the estimator model in Fig.8-b. Clearly, the area under the curves is very high for all the angle classes with a high true positive estimation rate and a low false positive estimation rate.



(a) (b)

Figure 8: ANN estimator model. (a) Confusion matrix, (b) ROC Curves.

Fig.9 shows the confusion matrix that describes the performance of the mapping table head orientation estimator model. This estimator model also has a very high accuracy (97.2%) and very high sensitivity and precision. However, this estimator model perform only slow estimations (only when there is high SNR ratio). Furthermore, this model can only estimate three classes of angles but on the other hand it requires less computation complexity compared to the ANN model.

Figure 9: Mapping table estimator model confusion matrix.

Finally, we evaluated the performance of the equalizer filter. Fig.10 shows the PSD for a speech signal with -90° (to the left) head orientation before (red) and after (black) applying the corresponding equalizer filter (blue). The green PSD is for a reference speech signal which we expect the equalized speech signal PSD be the same. As it can be seen, the designed filter works very good boosting the power at the desired frequencies and hence compensates the selective frequency power attenuation resulting from that head turning.

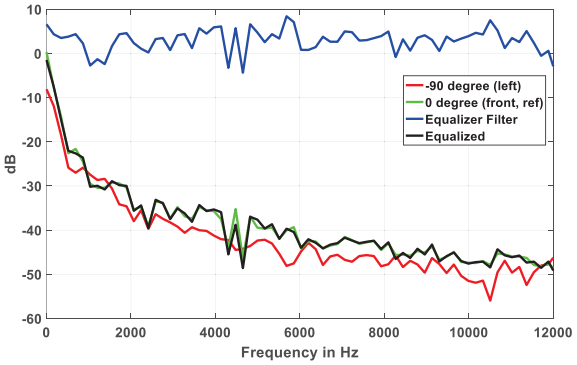


Figure 10: Speech signal enhancement using equalizer filter.

5. Conclusion

In this paper, we proposed two acoustic head orientation estimation models and the corresponding speech enhancement methods for ICC systems. Both approaches have been tested and evaluated under various SNR conditions and both provides promising results. However, on one hand, estimating the head orientation using an ANN model provides high estimation accuracy but in exchange of high memory and computational complexity requirements. On the other hand, inexpensive accurate head orientation estimation can be achieved using a simple mapping table model which maps the power ratio measures into a corresponding head orientation angle, but in exchange that high SNR situations are required to perform the estimation and fewer classes of head orientation angles can be estimated. However, the mapping table approach can become an interesting alternative for instance when memory and computational cost are constrained. Furthermore, we implemented equalization filters and noise reduction scheme and evaluated their performance. The next step is to deploy and investigate the performance of our methods with real time processing.

References

[1] Brian B. Monsona and Eric J. Hunter: Horizontal directivity of low- and high-frequency energy, J. Acoust. Soc. Am. 132 (1), 2012 Acoustical Society of America.

[2] A.Y. Nakano, S. Nakagawa, K. Yamamoto, Distant speech recognition using a microphone array network, IEICE Trans. Inf. Syst. E93-D (9) (2010) 2451–2462.

[3] S. Hwang, Y. Park, Y. Park, Sound direction estimation using an artificial ear for robots, Robot. Auton. Syst. 59 (3–4) (2011) 208–217.

[4] S. T. Shivappa, B.D. Rao, M.M. Trivedi, Role of head pose estimation in speech acquisition from distant microphones, in: Proceedings of ICASSP, 2009, pp. 3557–3560.

[5] Rasool Al-Mafrachi, Marco Gimm, Gerhard Schmidt. Acoustic estimation of the head orientation for In-Car Communication systems, in: Proceedings of DAGA München, 2018, pp. 1780–1783.

[6] C. Baasch: Verbesserung und Implementierung einer Geräuschschätzung in einem Echtzeitsystem für Anwendungen im Automobilbereich, Bachelor thesis, Kiel University, Faculty of Engineering, 2012. (In German).